Bernoulli **27**(4), 2021, 2267–2299
https://doi.org/10.3150/20-BEJ1308

The linear conditional expectation in Hilbert space

ILJA KLEBANOV[1,*], BJÖRN SPRUNGK[2] and T.J. SULLIVAN[1,3,†]

[1]*Zuse Institute Berlin, Takustraße 7, 14195 Berlin, Germany. E-mail:* [*]*klebanov@zib.de;* [†]*sullivan@zib.de*
[2]*Technische Universität Bergakademie Freiberg, 09596 Freiberg, Germany.*
E-mail: bjoern.sprungk@math.tu-freiberg.de
[3]*Mathematics Institute and School of Engineering, The University of Warwick, Coventry, CV4 7AL, United Kingdom. E-mail: t.j.sullivan@warwick.ac.uk*

The *linear conditional expectation* (LCE) provides a best linear (or rather, affine) estimate of the conditional expectation and hence plays an important rôle in approximate Bayesian inference, especially the *Bayes linear* approach. This article establishes the analytical properties of the LCE in an infinite-dimensional Hilbert space context. In addition, working in the space of affine Hilbert–Schmidt operators, we establish a regularisation procedure for this LCE. As an important application, we obtain a simple alternative derivation and intuitive justification of the *conditional mean embedding* formula, a concept widely used in machine learning to perform the conditioning of random variables by embedding them into reproducing kernel Hilbert spaces.

Keywords: Bayes linear analysis; conditional mean embedding; reproducing kernel Hilbert space; linear conditional expectation

1. Introduction

The crucial step in most inference problems is the approximation of the conditional expectation $\mathbb{E}[U|V]$, where $U \in L^2(\Omega, \Sigma, \mathbb{P}; \mathcal{G})$ and $V \in L^2(\Omega, \Sigma, \mathbb{P}; \mathcal{H})$ are random variables over some probability space $(\Omega, \Sigma, \mathbb{P})$ taking values in some separable Hilbert spaces $\mathcal{G}$ and $\mathcal{H}$, respectively. In Bayesian statistics, where it relates to the posterior mean, $\mathbb{E}[U|V]$ is an important point estimator of the inferred parameter. It is well known[1] that $\mathbb{E}[U|V]$ is the best approximation of U by a $\sigma(V)$-measurable random variable within $L^2(\Omega, \sigma(V), \mathbb{P}; \mathcal{G})$ (i.e. the orthogonal projection of U onto $L^2(\Omega, \sigma(V), \mathbb{P}; \mathcal{G})$),

$$\mathbb{E}[U|V] = \operatorname*{arg\,min}_{\tilde{U} \in L^2(\Omega, \sigma(V); \mathcal{G})} \|U - \tilde{U}\|_{L^2(\Omega, \Sigma, \mathbb{P}; \mathcal{G})} = \operatorname*{arg\,min}_{\tilde{U} \in L^2(\Omega, \sigma(V); \mathcal{G})} \mathbb{E}\big[\|U - \tilde{U}\|_{\mathcal{G}}^2\big]. \tag{1.1}$$

By the Doob–Dynkin representation [21], Lemma 1.13, the conditional expectation can therefore be rewritten in the form

$$\mathbb{E}[U|V] = \gamma_{U|V} \circ V \quad \mathbb{P}\text{-almost surely}, \tag{1.2}$$

where $\gamma_{U|V} : \mathcal{H} \to \mathcal{G}$ is a measurable map which we will call the *conditional expectation function* (CEF). In the language of statistical learning theory (or statistical decision theory), $\gamma_{U|V}$ is called the *regression function* and constitutes a Bayes predictor for the least squares error loss, that is, the predictor with the minimal risk [20], Section 2.4, which follows directly from (1.1).

[1]For $\mathbb{R}$-valued random variables see, for example, [10], Theorem 10.2.9; the general case follows by choosing orthonormal bases.

While computing $\gamma_{U|V}$, which is the main object of interest, is infeasible in most applications, various estimates can be constructed. The most prominent approach is to approximate $\gamma_{U|V}$ within the class $\mathsf{A}(\mathcal{H};\mathcal{G})$ of bounded affine operators[2] from $\mathcal{H}$ to $\mathcal{G}$, since this provides an explicit formula for the *linear conditional expectation function* (LCEF) $\gamma^{\mathsf{A}}_{U|V}$ under appropriate conditions [12], Lemma 4.1:

$$\gamma^{\mathsf{A}}_{U|V}(v) = \mu_U + C_{UV}C_V^{-1}(v - \mu_V), \tag{1.3}$$

where μ_U and μ_V denote the means and C_{UV} and C_V denote the cross-covariance and covariance operators of U and V, as defined in Section 3.

While the *linear conditional expectation* (LCE) $\mathbb{E}^{\mathsf{A}}[U|V] := \gamma^{\mathsf{A}}_{U|V} \circ V$ (also known as *Bayes linear estimator* or *adjusted expectation*) has been discussed extensively by Michael Goldstein and his collaborators in the framework of Bayes linear statistics mostly from an application point of view [18], a rigorous mathematical analysis of the LCE is yet to be established, especially for the case of infinite-dimensional $\mathcal{G}$ and $\mathcal{H}$. This level of generality, which this article seeks to provide, yields not just a satisfying mathematical theory but is also necessary for the application of LCE-type methods to problems with high-dimensional unknowns or data, such as time series and functional data analysis.

The first contribution of this paper is to fill this theoretical gap by studying the properties of the LCE and generalising formula (1.3) to infinite-dimensional Hilbert spaces. Thus far, (1.3) has been derived under the assumptions that $\mathcal{H}$ is finite-dimensional and that C_V is invertible [12], Section 4. In addition, working in the spaces of (affine) Hilbert–Schmidt operators, we establish a rigorous justification for the regularised version of (1.3).

Our second contribution is a simple alternative derivation and intuitive explanation of the widely used formula for the conditional mean embedding (CME), a method used in machine learning to perform the conditioning of random variables by embedding them into RKHSs, where it reduces to an affine transformation similar to (1.3) [16,22,33]. This result follows almost directly from the fact that, by the reproducing property, $\mathbb{E}[U|V]$ coincides with its best affine approximation $\mathbb{E}^{\mathsf{A}}[U|V]$.

Note that this paper considers only centered (cross-)covariance operators defined by (3.1). Some, but not all, of the results can be proven similarly for uncentered operators defined by (3.2), the theory for which is less general, since it allows only for strictly linear instead of affine approximations, that is, one would be restricted to fitting the probability density or data points in Figure 1 with a straight line through the origin.

This paper is structured as follows. Section 2 briefly surveys related work in statistics, machine learning, and dynamical systems. Section 3 establishes notation and standing assumptions for the remainder of the paper. Section 4 forms the core of the paper, in which we study the rigorous generalisation of the LCE to the infinite-dimensional Hilbert space context and also consider multiple formulations of the linear conditional covariance operator. We analyse their basic properties (Theorems 4.5 and 4.7) and derive explicit formulae for them in several regimes (Theorems 4.8, 4.13, and 4.14). In Sections 5 and 6, these ideas are applied to kernel conditional mean embeddings of random variables into RKHSs and to the conditioning of infinite-dimensional Gaussian random vectors, respectively. Some closing remarks are given in Section 7, after which all the proofs of results in the main text are given in Section 8. The Appendix contains technical supporting results.

[2]We note here an unfortunate but seemingly unavoidable clash of terminology: while the approximate conditional expectation (1.3) is usually called the *linear* conditional expectation in the literature, it in fact corresponds to approximation using *affine* operators.

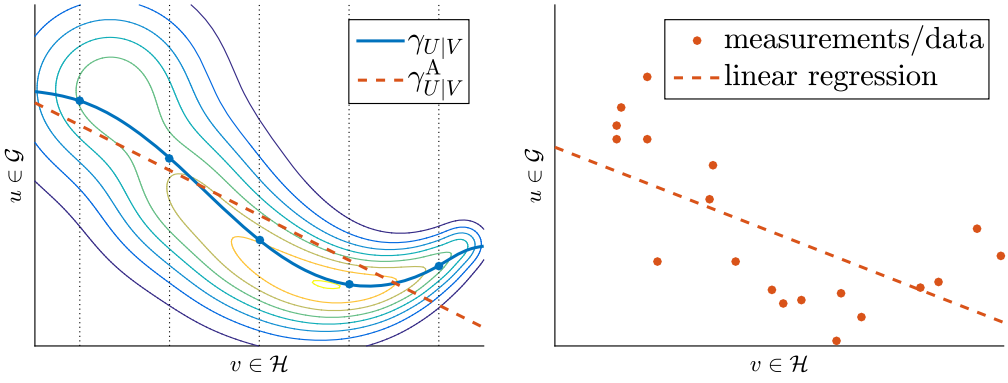

Figure 1. *Left*: Comparison of the conditional expectation function (CEF) $\gamma_{U|V} : \mathcal{H} \to \mathcal{G}$ and the linear conditional expectation function (LCEF) $\gamma^{\mathsf{A}}_{U|V} \in \mathsf{A}(\mathcal{H}; \mathcal{G})$. The contour plot shows the probability density ρ_{VU} of (V, U). *Right*: For an empirical probability distribution (e.g., given by data), the LCEF coincides with the solution to the linear least squares regression problem.

2. Related work

Formula (1.3) is the fundamental solution in linear least squares regression (or general linear models) and can be interpreted as the *best linear unbiased estimate* (BLUE); see, for example, [20], Sections 2.3.1 and 3.2. Figure 1 illustrates the connection between $\gamma^{\mathsf{A}}_{U|V}$ and linear regression: the two coincide if the probability distribution $\mathbb{P}_{VU}$ of (V, U) is an empirical distribution $\mathbb{P}_{VU} = J^{-1} \sum_{j=1}^{J} \delta_{(v_j, u_j)}$, where (v_j, u_j), $j = 1, \ldots, J$, are (or can be thought of as) measurements or data points.

Apart from the connection to linear regression, this work is related to several fields of applied mathematics. First and foremost, it should be seen as a systematic and rigorous treatment as well as an extension of *Bayes linear analysis*, which has been introduced and investigated by Michael Goldstein and his collaborators, see, for example, [17] and [18] and the references therein. Furthermore, [34], page 9, offers a Bayesian interpretation of the BLUE in special cases, namely that in the "uninformative" infinite-variance limit of a Gaussian prior, the limiting posterior is Gaussian with the BLUE as its conditional expectation.

The LCE is applied in a variety of fields:

In geostatistics, the LCE appears in form of the Kriging estimate for the value of a random field at unexplored locations given available (noisy) data of the random field at measurement locations [6,34].

In data assimilation, formula (1.3) defines the update scheme of the Kálmán filter and its many variants, including the ensemble Kálmán filter [13]. Although this update rule is typically interpreted as a Gaussian approximation – "in the large ensemble size limit the EnKF [...] does not reproduce the filtering distribution, except in the linear Gaussian case" [31] – it has been argued by [12], Section 4, that it should rather be seen as the best linear approximation of the required conditional expectations.

In machine learning, the method of *conditional mean embedding* (CME; [16,33]) applies the conditioning formula (1.3) to random variables embedded into RKHSs, where it becomes exact (i.e., $\gamma^{\mathsf{A}}_{U|V} = \gamma_{U|V}$) under certain conditions; see [22]. Section 5 provides an alternative derivation of the CME formula based on linear conditional expectations and thereby a natural justification of CMEs based in BLUEs; to the best of our knowledge, this connection has not been made before.

In the field of dynamical systems, the LCE is an important estimate of the *Koopman operator*

$$\mathcal{K}_\tau \colon L^\infty(\mathcal{X}) \to L^\infty(\mathcal{X}), \qquad \mathcal{K}_\tau f(x) := \mathbb{E}\big[f(X_{t+\tau}) | X_t = x\big],$$

of a time-homogeneous Markov chain $(X_t)_{t\in\mathcal{T}}$, where $\tau > 0$, $\mathcal{X} \subseteq \mathbb{R}^d$, and $\mathcal{T}$ is typically $\mathbb{Z}$, $\mathbb{N}$, $\mathbb{R}$, or $[0, \infty)$. Independently of one another, the dynamical systems, fluid dynamics, and molecular dynamics communities have developed data-driven dimensionality reduction techniques based on the eigendecomposition of this approximation, resulting in the methods of *time-lagged independent component analysis* (TICA) and *dynamic mode decomposition* (DMD). More generally, the data can be transformed to some feature space $\tilde{\mathcal{X}}$ by a feature map $\psi\colon \mathcal{X} \to \tilde{\mathcal{X}}$ prior to computing the linear approximation, resulting in the *variational approach of conformation dynamics* (VAC) or *empirical dynamic mode decomposition* (EDMD). The connections among these methods have been discussed in detail by [25]; see also the references therein to the original papers on these methods. As in the case of CMEs, if the feature map ψ is the canonical feature map $\psi(x) = k(x, \bullet)$ of an RKHS $\mathcal{H}$ with reproducing kernel k, then our analysis is applicable and it reveals the exactness of conditioning formula (1.3), thereby annihilating one of the main sources of error – the other being the approximation of the means and covariance operators μ_U, μ_V, C_V, and C_{UV}. The resulting kernelised versions of VAC and EDMD have been studied by [32] and [26]; the exactness of (1.3), which we establish, strengthens the analytical power of these methods. We also mention that there are time-inhomogeneous variants of these methods, namely *coherent mode decomposition* (CMD) and the *variational approach for Markov processes* (VAMP) and their kernelised version *kernel canonical correlation analysis* (kernel CCA; [24]); again, the established exactness of formula (1.3) in RKHSs can be exploited for analysing these approaches.

3. Preliminaries and notation

This paper will make use of various standing assumptions and items of notation, which we collect here for easy reference.

Throughout, $(\mathcal{F}, \langle\bullet, \bullet\rangle_{\mathcal{F}})$, $(\mathcal{G}, \langle\bullet, \bullet\rangle_{\mathcal{G}})$, and $(\mathcal{H}, \langle\bullet, \bullet\rangle_{\mathcal{H}})$ will be separable real Hilbert spaces and $U \in L^2(\Omega, \Sigma, \mathbb{P}; \mathcal{G})$, $V \in L^2(\Omega, \Sigma, \mathbb{P}; \mathcal{H})$, and $W \in L^2(\Omega, \Sigma, \mathbb{P}; \mathcal{F})$, where $(\Omega, \Sigma, \mathbb{P})$ is a fixed probability space.

$\mathcal{G}$-valued expected values $\mu_U := \mathbb{E}[U] := \int_\Omega U(\omega)\,\mathrm{d}\mathbb{P}(\omega)$ are always meant in the sense of a Bochner integral [8], Chapter II, as are the cross-covariance operators

$$C_{UV} := \mathbb{C}\mathrm{ov}[U, V] := \mathbb{E}\big[\big(U - \mathbb{E}[U]\big) \otimes \big(V - \mathbb{E}[V]\big)\big] = \mathbb{E}[U \otimes V] - \mathbb{E}[U] \otimes \mathbb{E}[V] \tag{3.1}$$

from $\mathcal{H}$ into $\mathcal{G}$, where, for $h \in \mathcal{H}$ and $g \in \mathcal{G}$, the outer product $g \otimes h\colon \mathcal{H} \to \mathcal{G}$ is the rank-one linear operator $(g \otimes h)(h') := \langle h, h'\rangle_{\mathcal{H}} g$. Naturally, we write $C_U = \mathbb{C}\mathrm{ov}[U]$ for the covariance operator $\mathbb{C}\mathrm{ov}[U, U]$, which is self-adjoint, non-negative, and trace-class [3,30], and all of the above reduces to the usual definitions in the scalar-valued case. Using Theorem A.3 and [28], Lemmas 16.7 and 16.21, it follows that the cross-covariance operators C_{UV} and $C_{VU} = C_{UV}^*$ are also trace-class (and in particular Hilbert–Schmidt) operators. In Section 5, we will briefly consider *uncentred* (cross-)covariance operators

$${}^{u}C_{UV} := {}^{u}\mathbb{C}\mathrm{ov}[U, V] := \mathbb{E}[U \otimes V], \qquad {}^{u}C_U := {}^{u}\mathbb{C}\mathrm{ov}[U] := {}^{u}\mathbb{C}\mathrm{ov}[U, U]. \tag{3.2}$$

The orthogonal projection onto a closed linear subspace F of a Hilbert space $\mathcal{H}$ will be denoted by $P_F^{\mathcal{H}}$, or just P_F whenever $\mathcal{H}$ is clear from context. Further, we abbreviate $L^2(\Omega, \Sigma, \mathbb{P}; \mathcal{G})$ by $L^2(\mathbb{P}; \mathcal{G})$ and further by $L^2(\mathbb{P})$ if $\mathcal{G} = \mathbb{R}$; and $L^2(\Omega, \tilde{\Sigma}, \mathbb{P}|_{\tilde{\Sigma}}; \mathcal{G})$ by $L^2(\Omega, \tilde{\Sigma}; \mathcal{G})$ for any sub-σ-algebra $\tilde{\Sigma} \subseteq \Sigma$. $\mathbb{P}_X$ denotes the distribution of a random variable $X\colon \Omega \to \mathcal{X}$, that is, the pushforward $X_\#\mathbb{P}$ of $\mathbb{P}$ under X.

For a linear operator $A\colon \mathcal{H} \to \mathcal{G}$ between Hilbert spaces $\mathcal{H}$ and $\mathcal{G}$, its Moore–Penrose pseudo-inverse $A^\dagger\colon \operatorname{dom} A^\dagger \to \mathcal{H}$ is the unique extension of

$$A|^{-1}_{(\ker A)^\perp}\colon \operatorname{ran} A \to (\ker A)^\perp$$

to a linear operator $A^\dagger$ defined on $\operatorname{dom} A^\dagger := (\operatorname{ran} A) \oplus (\operatorname{ran} A)^\perp \subseteq \mathcal{G}$ subject to the criterion that $\ker A^\dagger = (\operatorname{ran} A)^\perp$. In general, $\operatorname{dom} A^\dagger$ is a dense but proper subpace of $\mathcal{G}$ and $A^\dagger$ is an unbounded operator; global definition and boundedness of $A^\dagger$ occur precisely when $\operatorname{ran} A$ is closed in $\mathcal{G}$ [11], Section 2.1.

The following spaces of linear and affine operators from $\mathcal{H}$ to $\mathcal{G}$ will play a fundamental rôle in the approximation of $\gamma_{U|V}$.

Definition 3.1. Let $\mathcal{H}$, $\mathcal{G}$, and V be as above. We define the following spaces of linear and affine operators from $\mathcal{H}$ to $\mathcal{G}$:

$$\begin{aligned}
\mathsf{L}(\mathcal{H};\mathcal{G}) &:= \{\gamma\colon \mathcal{H}\to\mathcal{G} \mid \gamma \text{ is a bounded linear operator}\},\\
\mathsf{A}(\mathcal{H};\mathcal{G}) &:= \big\{\gamma\colon \mathcal{H}\to\mathcal{G} \mid \gamma(h) = b + Ah \text{ for some } b\in\mathcal{G},\, A\in\mathsf{L}(\mathcal{H};\mathcal{G})\big\},\\
\mathsf{L}_V(\mathcal{H};\mathcal{G}) &:= \big\{\gamma\colon \mathcal{H}\to\mathcal{G} \mid \gamma \text{ is linear and } \gamma\circ V\in L^2(\mathbb{P};\mathcal{G})\big\},\\
\mathsf{A}_V(\mathcal{H};\mathcal{G}) &:= \big\{\gamma\colon \mathcal{H}\to\mathcal{G} \mid \gamma(h) = b + Ah \text{ for some } b\in\mathcal{G},\, A\in\mathsf{L}_V(\mathcal{H};\mathcal{G})\big\},\\
\mathsf{L}_2(\mathcal{H};\mathcal{G}) &:= \{\gamma\colon \mathcal{H}\to\mathcal{G} \mid \gamma \text{ is a Hilbert–Schmidt operator}\},\\
\mathsf{A}_2(\mathcal{H};\mathcal{G}) &:= \big\{\gamma\colon \mathcal{H}\to\mathcal{G} \mid \gamma(h) = b + Ah \text{ for some } b\in\mathcal{G},\, A\in\mathsf{L}_2(\mathcal{H};\mathcal{G})\big\}.
\end{aligned}$$

Note well that elements of $\mathsf{L}_V(\mathcal{H};\mathcal{G})$ and $\mathsf{A}_V(\mathcal{H};\mathcal{G})$ may be unbounded operators, although their unboundedness is in some sense restricted by the square-integrability requirement. For any collection Γ of affine or linear operators $\gamma\colon \mathcal{H}\to\mathcal{G}$, we set $\Gamma\circ V := \{\gamma\circ V \mid \gamma\in\Gamma\}$ and

$$\overline{\mathsf{L}_\mathcal{G}\circ V} := \overline{\mathsf{L}(\mathcal{H};\mathcal{G})\circ V}^{L^2(\Omega,\Sigma,\mathbb{P};\mathcal{G})}, \qquad \overline{\mathsf{A}_\mathcal{G}\circ V} := \overline{\mathsf{A}(\mathcal{H};\mathcal{G})\circ V}^{L^2(\Omega,\Sigma,\mathbb{P};\mathcal{G})}.$$

Here and henceforth, overlines and superscripts denote topological closures. The operator norm will be denoted by $\|\cdot\|$. For any affine operator $\gamma\in\mathsf{A}_V(\mathcal{H};\mathcal{G})$, $\gamma(h) = b + Ah$, $b\in\mathcal{G}$, $A\in\mathsf{L}_V(\mathcal{H};\mathcal{G})$, we define the "non-affine part" by $\overline{\gamma} := A$. The Hilbert–Schmidt inner product will be denoted by $\langle\gamma_1,\gamma_2\rangle_{\mathsf{L}_2} := \operatorname{tr}(\gamma_1^*\gamma_2) = \operatorname{tr}(\gamma_1\gamma_2^*)$, where $\gamma_1,\gamma_2\in\mathsf{L}_2(\mathcal{H};\mathcal{G})$, and the corresponding norm by $\|\cdot\|_{\mathsf{L}_2}$. Further, for $\gamma,\gamma'\in\mathsf{A}_2(\mathcal{H};\mathcal{G})$, we define the seminorm $\|\gamma\|_{\mathsf{A}_2} := \|\overline{\gamma}\|_{\mathsf{L}_2}$ and the semi-inner product $\langle\gamma,\gamma'\rangle_{\mathsf{A}_2} := \langle\overline{\gamma},\overline{\gamma}'\rangle_{\mathsf{L}_2}$.

Proposition 3.2. *With the notation above,*

$$\overline{\mathsf{L}_\mathcal{G}\circ V} \subseteq \mathsf{L}_V(\mathcal{H};\mathcal{G})\circ V, \qquad \overline{\mathsf{A}_\mathcal{G}\circ V} \subseteq \mathsf{A}_V(\mathcal{H};\mathcal{G})\circ V.$$

4. Linear conditional expectation and covariance

It is well known that the conditional expectation $\mathbb{E}[U|V]$ is the orthogonal projection of U onto $L^2(\Omega,\sigma(V),\mathbb{P};\mathcal{G})$; see Footnote 1. Since $\mathbb{E}[U|V]$ is $\sigma(V)$-measurable, the Doob–Dynkin lemma

[21], Lemma 1.13, implies the existence of a Borel-measurable function $\gamma_{U|V}\colon \mathcal{H} \to \mathcal{G}$ such that $\mathbb{E}[U|V] = \gamma_{U|V} \circ V$ a.s. In particular, $\gamma_{U|V}$ minimizes the functional

$$\mathcal{E}_{U|V}(\gamma) := \|U - \gamma \circ V\|^2_{L^2(\Omega,\Sigma,\mathbb{P};\mathcal{G})} = \mathbb{E}\big[\|U - \gamma \circ V\|^2_{\mathcal{G}}\big] \tag{4.1}$$

within the class of Borel-measurable functions $\gamma\colon \mathcal{H} \to \mathcal{G}$. Since $\mathbb{E}[U|V]$ is unique (as an orthogonal projection), $\gamma_{U|V}$ is unique $\mathbb{P}_V$-a.e. and we set $\mathbb{E}[U|V = v] := \gamma_{U|V}(v)$ for $v \in \mathcal{H}$.

It seems natural to define the best linear approximation (see Footnote 2) of the conditional expectation as $\mathbb{E}^{\mathsf{A}}[U|V] = \gamma^{\mathsf{A}}_{U|V} \circ V$, where $\gamma^{\mathsf{A}}_{U|V}$ minimizes $\mathcal{E}_{U|V}(\gamma)$ within the class $\mathsf{A}(\mathcal{H};\mathcal{G})$ of bounded affine operators, in other words, as the $L^2(\mathbb{P};\mathcal{G})$-orthogonal projection of U onto $\mathsf{A}(\mathcal{H};\mathcal{G}) \circ V$. Since this space is not closed in $L^2(\mathbb{P};\mathcal{G})$, the proper definition uses the projection onto its closure. In line with the definition of the conditional covariance operator

$$\mathbb{C}\mathrm{ov}[U, W|V] := \mathbb{E}\big[R[U|V] \otimes R[W|V] \,|\, V\big], \qquad R[U|V] := U - \mathbb{E}[U|V], \tag{4.2}$$

we further define the linear conditional covariance operator $\mathbb{C}\mathrm{ov}^{\mathsf{A}}[U, W|V]$ as follows.

Definition 4.1. With the notation of Section 3, define the *linear conditional expectation* (LCE) $\mathbb{E}^{\mathsf{A}}[U|V]$ (also called *adjusted expectation*, [18], Section 3.1), the *linear conditional residual* $R^{\mathsf{A}}[U|V]$, and the *linear conditional covariance operator* (LCC) $\mathbb{C}\mathrm{ov}^{\mathsf{A}}[U, W|V]$ of U given V by

$$\begin{aligned}
\mathbb{E}^{\mathsf{A}}[U|V] &:= P_{\overline{\mathsf{A}_{\mathcal{G}} \circ V}} U,\\
R^{\mathsf{A}}[U|V] &:= U - \mathbb{E}^{\mathsf{A}}[U|V],\\
\mathbb{C}\mathrm{ov}^{\mathsf{A}}[U, W|V] &:= \mathbb{E}^{\mathsf{A}}\big[R^{\mathsf{A}}[U|V] \otimes R^{\mathsf{A}}[W|V] \,|\, V\big].
\end{aligned}$$

Further, define the *average linear conditional covariance operator* (ALCC) by

$$\mathbb{C}\mathrm{ov}^{\mathsf{A}}_V[U, W] := \mathbb{E}\big[R^{\mathsf{A}}[U|V] \otimes R^{\mathsf{A}}[W|V]\big].$$

By Proposition 3.2, $\mathbb{E}^{\mathsf{A}}[U|V]$ will be of the form $\gamma^{\mathsf{A}}_{U|V} \circ V$, where $\gamma^{\mathsf{A}}_{U|V} \in \mathsf{A}_V(\mathcal{H};\mathcal{G})$ is unique $\mathbb{P}_V$-a.e. and will be referred to as the *linear conditional expectation function* (LCEF). As usual, we set $\mathbb{C}\mathrm{ov}[U|V] := \mathbb{C}\mathrm{ov}[U, U|V]$, $\mathbb{C}\mathrm{ov}^{\mathsf{A}}[U|V] := \mathbb{C}\mathrm{ov}^{\mathsf{A}}[U, U|V]$, and $\mathbb{C}\mathrm{ov}^{\mathsf{A}}_V[U] := \mathbb{C}\mathrm{ov}^{\mathsf{A}}_V[U, U]$.

Since $\mathbb{E}^{\mathsf{A}}[U|V]$ and $\mathbb{C}\mathrm{ov}^{\mathsf{A}}[U|V]$ are defined as $L^2(\mathbb{P};\mathcal{G})$-orthogonal projection, all statements and identities in the following subsections only hold $\mathbb{P}$-a.s.

Remark 4.2. [18], Section 3.3, call $\mathbb{C}\mathrm{ov}^{\mathsf{A}}_V[U, W]$ the *adjusted covariance* and argue that this is the proper way to define the linear analogue of the conditional covariance. However, this definition is not in line with the classical conditional covariance (4.2) because it fails to condition on V a second time; see Example 4.3 below. Note that the ALCC $\mathbb{C}\mathrm{ov}^{\mathsf{A}}_V[U, W]$ is therefore not a random variable but rather the expected value of the LCC $\mathbb{C}\mathrm{ov}^{\mathsf{A}}[U, W|V]$, hence our term "average linear conditional covariance"; see Theorem 4.15, where we also show that it coincides with the well-known Gaussian conditional covariance formula, $\mathbb{C}\mathrm{ov}^{\mathsf{A}}_V[U, W] = C_{UW} - C_{UV}C_V^{\dagger}C_{VW}$ (and similarly its more general version in the incompatible case). While $\mathbb{C}\mathrm{ov}^{\mathsf{A}}_V[U, U]$ is always non-negative (see Theorem 4.15(b)), $\mathbb{C}\mathrm{ov}^{\mathsf{A}}[U, W|V]$ can take on negative values (see Theorem 4.7(i)).

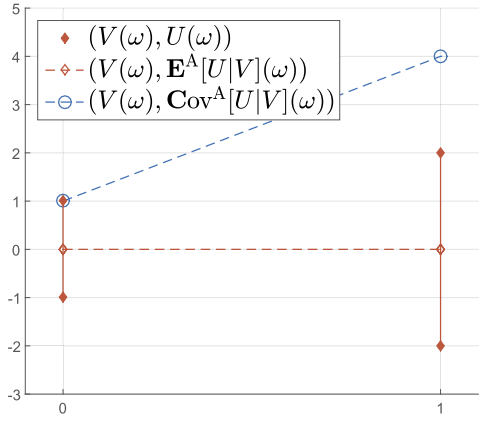

Figure 2. In Example 4.3, the conditional expectation $\mathbb{E}[U|V]$ as well as the conditional covariance $\mathbb{C}\mathrm{ov}[U|V]$ happen to be affine, which is why they coincide with the LCE $\mathbb{E}^{\mathsf{A}}[U|V]$ and the LCC $\mathbb{C}\mathrm{ov}^{\mathsf{A}}[U|V]$, respectively. The ALCC $\mathbb{C}\mathrm{ov}^{\mathsf{A}}_{V}[U] = \frac{5}{2}$ captures the expected value of $\mathbb{C}\mathrm{ov}^{\mathsf{A}}[U|V]$.

Example 4.3. Consider the following simple example of an LCE and an LCC. Let $\mathcal{H} = \mathcal{G} = \mathbb{R}$, let $\mathbb{P}$ be the uniform distribution on $\Omega = \{1, 2, 3, 4\}$, and let V and U be as defined below and illustrated in Figure 2.

$\omega \in \Omega$	$V(\omega)$	$U(\omega)$	$\mathbb{E}^{\mathsf{A}}[U\|V](\omega)$	$\mathbb{C}\mathrm{ov}^{\mathsf{A}}[U\|V](\omega)$
1	0	1	0	1
2	0	−1	0	1
3	1	2	0	4
4	1	−2	0	4

By symmetry, $\mathbb{E}^{\mathsf{A}}[U|V] = \mathbb{E}[U|V] = 0$ and, since $(V, R^{\mathsf{A}}[U|V]^2)$ takes on only two values, $\mathbb{C}\mathrm{ov}^{\mathsf{A}}[U|V]$ coincides with the classical conditional covariance $\mathbb{C}\mathrm{ov}[U|V]$. The ALCC $\mathbb{C}\mathrm{ov}^{\mathsf{A}}_{V}[U] = \frac{5}{2}$ only captures its expected value.

The main aim of this section is to investigate basic properties of, and provide explicit formulae for, the LCE and the LCC.

4.1. Basic properties of the LCE

This section highlights, in Theorems 4.5 and 4.7, respectively, the ways in which the LCE shares and lacks the key properties of the exact conditional expectation. We call attention to the non-trivial conditions that appear to be necessary for the LCE version of the dominated convergence theorem; see Theorem 4.5(i), Remark 4.6, and Theorem 4.7(g). As mentioned above, all statements concerning $\mathbb{E}^{\mathsf{A}}[U|V]$ and $\mathbb{C}\mathrm{ov}^{\mathsf{A}}[U|V]$ only hold $\mathbb{P}$-a.s.

Lemma 4.4. *With the notation of Section 3, let* $W \in \overline{\mathsf{A}_{\mathcal{F}} \circ V}$. *Then, a.s.,*

$$\mathbb{E}\big[R^{\mathsf{A}}[U|V]\big] = 0, \qquad \mathbb{C}\mathrm{ov}\big[R^{\mathsf{A}}[U|V], W\big] = 0.$$

In particular,

$$\mathbb{C}\mathrm{ov}\big[\mathbb{E}^{\mathsf{A}}[U|V], V\big] = \mathbb{C}\mathrm{ov}[U, V] \quad a.s.$$

By way of comparison with the exact conditional expectation, some basic properties of the LCE are summarised by the following theorem (a martingale property together with a martingale convergence theorem are postponed to Theorem 4.12).

Theorem 4.5 (Basic properties satisfied by the LCE and the LCC). *With the notation of Section 3, let U' and $U_k \in L^2(\Omega, \Sigma, \mathbb{P}; \mathcal{G})$ for $k \in \mathbb{N}$. Further, let $\varphi \in \mathsf{A}(\mathcal{H}; \mathcal{F})$. The LCE fulfills the following basic properties $\mathbb{P}$-a.s.:*

(a) stability:
$\mathbb{E}^{\mathsf{A}}[U|V] = \mathbb{E}[U|V]$, *if* $\mathbb{E}[U|V] \in \overline{\mathsf{A}_{\mathcal{G}} \circ V}$, *in particular,*
$\mathbb{E}^{\mathsf{A}}[U|V] = g$ *a.s. whenever* $U = g \in \mathcal{G}$ *a.s.*, $\mathbb{E}^{\mathsf{A}}[V|V] = V$ *and* $\mathbb{E}^{\mathsf{A}}[\varphi \circ V|V] = \varphi \circ V$;

(b) linearity:
$\mathbb{E}^{\mathsf{A}}[aU + bU'|V] = a\mathbb{E}^{\mathsf{A}}[U|V] + b\mathbb{E}^{\mathsf{A}}[U'|V]$ *for any* $a, b \in \mathbb{R}$ *and*
$\mathbb{E}^{\mathsf{A}}[\psi(U)|V] = \psi(\mathbb{E}^{\mathsf{A}}[U|V])$ *for any* $\psi \in \mathsf{A}(\mathcal{G}; \mathcal{F})$;

(c) self-adjointness:
$\mathbb{E}[\langle U', \mathbb{E}^{\mathsf{A}}[U|V]\rangle_{\mathcal{G}}] = \mathbb{E}[\langle \mathbb{E}^{\mathsf{A}}[U'|V], \mathbb{E}^{\mathsf{A}}[U|V]\rangle_{\mathcal{G}}] = \mathbb{E}[\langle \mathbb{E}^{\mathsf{A}}[U'|V], U\rangle_{\mathcal{G}}]$;

(d) law of total linear expectation:
$\mathbb{E}[\mathbb{E}^{\mathsf{A}}[U|V]] = \mathbb{E}[U]$;

(e) compatibility with conditional expectation:
$\mathbb{E}[\mathbb{E}^{\mathsf{A}}[U|V] \mid W] = \mathbb{E}^{\mathsf{A}}[\mathbb{E}[U|W] \mid V]$;
$\mathbb{E}[\mathbb{E}^{\mathsf{A}}[U|V] \mid V] = \mathbb{E}^{\mathsf{A}}[\mathbb{E}[U|V] \mid V] = \mathbb{E}^{\mathsf{A}}[U|V]$;

(f) tower properties:
$\mathbb{E}^{\mathsf{A}}[\mathbb{E}[U|V] \mid W] = \mathbb{E}^{\mathsf{A}}[U|W]$ *if* $\sigma(W) \subseteq \sigma(V)$;
$\mathbb{E}^{\mathsf{A}}[U \mid \mathbb{E}^{\mathsf{A}}[U|V]] = \mathbb{E}^{\mathsf{A}}[U|V]$;
$\mathbb{E}^{\mathsf{A}}[\mathbb{E}^{\mathsf{A}}[U|V] \mid \varphi \circ V] = \mathbb{E}^{\mathsf{A}}[\mathbb{E}[U|V] \mid \varphi \circ V] = \mathbb{E}^{\mathsf{A}}[U|\varphi \circ V]$, *in particular,*
$\mathbb{E}^{\mathsf{A}}[\mathbb{E}^{\mathsf{A}}[U|(V, W)] \mid V] = \mathbb{E}^{\mathsf{A}}[\mathbb{E}[U|(V, W)] \mid V] = \mathbb{E}^{\mathsf{A}}[U|V]$;

(g) law of total linear covariance:
$\mathbb{C}\mathrm{ov}[U, W] = \mathbb{C}\mathrm{ov}[\mathbb{E}^{\mathsf{A}}[U|V], \mathbb{E}^{\mathsf{A}}[W|V]] + \mathbb{E}[\mathbb{C}\mathrm{ov}^{\mathsf{A}}[U, W|V]]$, *in particular,*
$\mathbb{C}\mathrm{ov}[U] \geq \mathbb{C}\mathrm{ov}[\mathbb{E}[U|V]] \geq \mathbb{C}\mathrm{ov}[\mathbb{E}^{\mathsf{A}}[U|V]] \geq 0$.

(h) pulling out independent factors:
$\mathbb{E}^{\mathsf{A}}[W \otimes U|V] = \mathbb{E}[W] \otimes \mathbb{E}^{\mathsf{A}}[U|V]$, *if* W *is independent of* (U, V), *in particular,*
$\mathbb{E}^{\mathsf{A}}[W|V] = \mathbb{E}[W]$, *if* V *and* W *are independent* (*this also follows from* (a));

(i) L^2-dominated convergence theorem (DCT):
$\|\mathbb{E}^{\mathsf{A}}[U_k|V] - \mathbb{E}^{\mathsf{A}}[U|V]\|_{\mathcal{G}} \xrightarrow[k\to\infty]{\text{a.s.}} 0$ *if* C_V *has finite rank and either of the following conditions holds*:
(α) $\|U_k - U\|_{\mathcal{G}} \xrightarrow[k\to\infty]{\text{a.s.}} 0$ *and* $\|U_k\|_{\mathcal{G}} \leq Y$ *for all* $k \in \mathbb{N}$ *and some* $Y \in L^2(\mathbb{P})$,
(β) $\|U_k - U\|_{L^2(\mathbb{P};\mathcal{G})} \xrightarrow[k\to\infty]{} 0$.

Remark 4.6 (Sufficient conditions for the DCT). Note that in Theorem 4.5(i)(α) the dominating random variable $Y \in L^2(\mathbb{P})$ is assumed to be square-integrable, which is slightly stronger than the conventional assumption $Y \in L^1(\mathbb{P})$ (a counterexample to the sufficiency of the latter assumption is provided in Theorem 4.7(g)). On the other hand, (β) is a particularly weak assumption, too weak for analogous statements on the (regular) conditional expectation $\mathbb{E}[\bullet|V]$ in place of $\mathbb{E}^{\mathsf{A}}[\bullet|V]$.

So far, we could only prove the DCT under the condition that C_V has finite rank. Note that counterexamples can easily be constructed (see below) if one only assumes (β). However, the validity of the DCT under (α) for C_V of infinite rank remains an open problem. One obstacle here is the missing monotonicity of the LCE, see Theorem 4.7(a): $\|U_k\|_{\mathcal{G}} \le Y$ a.s. does not imply that $\|\mathbb{E}^{\mathsf{A}}[U_k|V]\|_{\mathcal{G}} \le Y$ a.s.

For a counterexample to the DCT under assumption (β) (without the finite-rank assumption) consider a centered Gaussian random variable V on $\mathcal{H} = \ell^2$ with Karhunen–Loève expansion $V = \sum_n \sigma_n Z_n e_n$, where $\sigma_n > 0$ for all $n \in \mathbb{N}$, $\sum_n \sigma_n^2 < \infty$, $Z_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ and $(e_n)_{n\in\mathbb{N}}$ is the canonical basis of $\mathcal{H} = \ell^2$. Choose $\mathcal{G} = \mathbb{R}$ and $U_k = \delta_k Z_k$ where $\delta_k \searrow 0$ such that $\mathbb{P}[A_k] = 1/k$ for $A_k = [\delta_k Z_k \ge \varepsilon]$ and $\varepsilon = 1$. Then $\mathbb{E}^{\mathsf{A}}[U_k|V] = U_k$ by definition of the LCE and, since the family of events $(A_k)_{k\in\mathbb{N}}$ is independent and $\sum_k \mathbb{P}[A_k] = \infty$, the second Borel–Cantelli lemma implies $\mathbb{P}[A_k \text{ i.o.}] = 1$. Hence, by [4], Theorem 4.1.6, $\mathbb{E}^{\mathsf{A}}[U_k|V] = U_k$ does not converge to zero a.s., while (β) is satisfied since $\delta_k \to 0$.

It is also worth mentioning that the LCE does *not* fulfill several important properties of conditional expectations. These are summarised by the following statements and the actual counterexamples are provided in the proof in Section 8. Note in particular that there are scalar-valued counterexamples in each case, and so these deficiencies of the LCE are not merely a consequence of the Hilbert space context.

Theorem 4.7 (Basic properties *not* satisfied by the LCE and the LCC). *For each of the following desired properties of the LCE, there is an explicit counterexample using $\mathbb{R}$-valued random variables U, U', V, etc. satisfying the assumptions of Theorem 4.5. As usual, all statements have to be understood $\mathbb{P}$-a.s.*

(a) monotonicity:$^{(\text{invalid})}$
$U \ge U'$ *implies* $\mathbb{E}^{\mathsf{A}}[U|V] \ge \mathbb{E}^{\mathsf{A}}[U'|V]$ *in the case* $\mathcal{G} = \mathbb{R}$;

(b) triangle inequality:$^{(\text{invalid})}$
$\|\mathbb{E}^{\mathsf{A}}[U|V]\|_{\mathcal{G}} \le \mathbb{E}^{\mathsf{A}}[\|U\|_{\mathcal{G}}|V]$;

(c) Jensen's inequality:$^{(\text{invalid})}$
$f(\mathbb{E}^{\mathsf{A}}[U|V]) \le \mathbb{E}^{\mathsf{A}}[f(U)|V]$ *for any convex*[3] *function* $f : \mathcal{G} \to \mathbb{R}$;

(d) pulling out known factors:$^{(\text{invalid})}$
$\mathbb{E}^{\mathsf{A}}[f(V)U|V] = f(V)\mathbb{E}^{\mathsf{A}}[U|V]$ *for measurable*[4] *maps* $f : \mathcal{H} \to \mathbb{R}$;

(e) (yet another) tower property:$^{(\text{invalid})}$
$\mathbb{E}^{\mathsf{A}}[\mathbb{E}^{\mathsf{A}}[U|V] \mid W] = \mathbb{E}^{\mathsf{A}}[U|W]$ *if* $\sigma(W) \subseteq \sigma(V)$;

(f) Fatou's lemma:$^{(\text{invalid})}$
Let $\mathcal{G} = \mathbb{R}$ *and* $\mathbb{E}^{\mathsf{A}}[\inf_{k\in\mathbb{N}} U_k|V] < \infty$ (*alternatively*, $\mathbb{E}[\inf_{k\in\mathbb{N}} U_k|V] < \infty$). *Then*
$\mathbb{E}^{\mathsf{A}}[\liminf_{k\to\infty} U_k \mid V] \le \liminf_{k\to\infty} \mathbb{E}^{\mathsf{A}}[U_k|V]$;

(g) L^1-dominated convergence theorem:$^{(\text{invalid})}$
If $\|U_k - U\|_{\mathcal{G}} \xrightarrow[k\to\infty]{\text{a.s.}} 0$ *and* $\|U_k\|_{\mathcal{G}} \le Y$ *for all* $k \in \mathbb{N}$ *and some* $Y \in L^1(\mathbb{P})$, *then* $\|\mathbb{E}^{\mathsf{A}}[U_k|V] - \mathbb{E}^{\mathsf{A}}[U|V]\|_{\mathcal{G}} \xrightarrow[k\to\infty]{\text{a.s.}} 0$.

(h) L^p-contractivity for $p \ne 2$:$^{(\text{invalid})}$
$\mathbb{E}^{\mathsf{A}}[\cdot|V]$ *is a contractive projection of* $L^p(\mathbb{P};\mathcal{G})$ *spaces for some* $1 \le p \ne 2$, *i.e.*
$\mathbb{E}[\|\mathbb{E}^{\mathsf{A}}[U|V]\|_{\mathcal{G}}^p] \le \mathbb{E}[\|U\|_{\mathcal{G}}^p]$ *for any* $U \in L^p(\Omega, \Sigma, \mathbb{P}; \mathcal{G})$.

[3] Note that Jensen's (in-)equality holds for affine functions $f \in \mathsf{A}(\mathcal{G}; \mathcal{F})$, cf. Theorem 4.5(b).
[4] Our counterexample shows that property (d) is invalid even if "measurable" is strengthened to "linear".

(i) non-negativity of the LCC:(invalid)
The LCC $\mathbb{C}\mathrm{ov}^{\mathsf{A}}[U|V]$ *is non-negative*, $\mathbb{C}\mathrm{ov}^{\mathsf{A}}[U|V] \geq 0$.

Note that, in contrast to Theorem 4.7(h), $\mathbb{E}^{\mathsf{A}}[\bullet|V]$ *is* a contractive projection on $L^2(\mathbb{P};\mathcal{G})$; this follows directly from the definition of the LCE as an $L^2(\mathbb{P};\mathcal{G})$-orthogonal projection.

4.2. Explicit formula for the LCE: Compatible case

We are first going assume that $\operatorname{ran} C_{VU} \subseteq \operatorname{ran} C_V$, which, following [7] and [29], we call the *compatible case*. In this case, the orthogonal projection $\mathbb{E}^{\mathsf{A}}[U|V]$ of U onto $\overline{\mathsf{A}_{\mathcal{G}} \circ V}$ turns out to lie in $\mathsf{A}(\mathcal{H};\mathcal{G}) \circ V$, which is generally not closed in $L^2(\Omega, \Sigma, \mathbb{P};\mathcal{G})$. The following theorem provides an explicit formula for the (affine) conditional mean and generalises [12], Lemma 4.1.

Theorem 4.8 (Formula for the LCE: compatible case). *With the notation of Section 3, if the range inclusion* $\operatorname{ran} C_{VU} \subseteq \operatorname{ran} C_V$ *holds, then the operator* $C_V^{\dagger} C_{VU} \colon \mathcal{G} \to \mathcal{H}$ *is bounded and the operator* $\gamma^{\mathsf{A}}_{U|V} \in \mathsf{A}(\mathcal{H};\mathcal{G})$ *defined by*

$$\gamma^{\mathsf{A}}_{U|V}(v) := \mu_U + \left(C_V^{\dagger} C_{VU}\right)^*(v - \mu_V)$$

minimizes the functional $\mathcal{E}_{U|V}$ *given by* (4.1) *within* $\mathsf{A}(\mathcal{H};\mathcal{G})$. *In particular,* $\mathbb{E}^{\mathsf{A}}[U|V] = \gamma^{\mathsf{A}}_{U|V} \circ V$ *a.s., that is,* $\gamma^{\mathsf{A}}_{U|V}$ *is an LCEF.*

Theorem 4.8 is a genuine generalisation of the case $\dim \mathcal{H} < \infty$ for the following reason, which is a direct consequence of Theorem A.3.

Corollary 4.9. *With the notation of Section 3, the condition* $\operatorname{ran} C_{VU} \subseteq \operatorname{ran} C_V$ *is always fulfilled whenever* C_V *has closed range, and, in particular, if* $\mathcal{H}$ *is finite dimensional.*

4.3. Explicit formula for the LCE: Incompatible case

We are now going to treat the general case, in which the orthogonal projection $\mathbb{E}^{\mathsf{A}}[U|V]$ of U can not be expected to lie in $\mathsf{A}(\mathcal{H};\mathcal{G}) \circ V$. We are therefore going to approximate $\mathbb{E}^{\mathsf{A}}[U|V]$ by a sequence of bounded (in fact, even finite-rank) operators $\gamma^{(n)}_{U|V} \in \mathsf{A}(\mathcal{H};\mathcal{G})$ composed with V, where the convergence will hold in the $L^2(\mathbb{P};\mathcal{G})$ norm as well as a.s. This requires some additional notation.

Notation 4.10. Let $\dim \mathcal{H} = \infty$[5] and recall the notation of Section 3. Consider the eigendecomposition of the covariance operator C_V,

$$C_V = \sum_{n \in \mathbb{N}} \sigma_n h_n \otimes h_n, \quad \sigma_n^2 \geq 0,$$

[5]This assumption is not substantial and we make it merely for the sake of simplifying our notation. Note that the finite-dimensional case has been analysed in the previous subsection.

where $(h_n)_{n\in\mathbb{N}}$ is a complete orthonormal system of $\mathcal{H}$. Let $n \in \mathbb{N}$, $\mathcal{H}^{(n)} := \operatorname{span}\{h_1, \dots, h_n\}$, $V^{(n)} := P_{\mathcal{H}^{(n)}} V$ and

$$C := \begin{pmatrix} C_U & C_{UV} \\ C_{VU} & C_V \end{pmatrix}, \qquad C^{(n)} := P_{\mathcal{H}^{(n)}} C P_{\mathcal{H}^{(n)}} = \begin{pmatrix} C_U & C_{UV}^{(n)} \\ C_{VU}^{(n)} & C_V^{(n)} \end{pmatrix}.$$

Since $C_V^{(n)}$ has finite rank, Theorem A.3 yields $\operatorname{ran} C_{VU}^{(n)} \subseteq \operatorname{ran} C_V^{(n)}$ and we can define the operator $\gamma_{U|V}^{(n)} \in \mathsf{A}(\mathcal{H}; \mathcal{G})$ by

$$\gamma_{U|V}^{(n)}(v) := \mu_U + \big(C_V^{(n)\dagger} C_{VU}^{(n)}\big)^*(v - \mu_V).$$

Further, by Theorem A.3 and adopting the notation therein, the operator $M_{VU} := (C_V^{1/2})^\dagger C_{VU} = R_{VU} C_U^{1/2} \colon \mathcal{G} \to \mathcal{H}$ is well defined and bounded (in fact, it is even Hilbert–Schmidt).

Lemma 4.11. *With the notation of Section 3 and using Notation 4.10,*

$$\overline{\mathsf{A}_{\mathcal{G}} \circ V} \cap L^2\big(\Omega, \sigma\big(V^{(n)}\big); \mathcal{G}\big) = \overline{\mathsf{A}_{\mathcal{G}} \circ V^{(n)}}. \tag{4.3}$$

Theorem 4.12 (Martingale property and martingale convergence theorem). *With the notation of Section 3 and using Notation 4.10,*

(a) $(\mathbb{E}^{\mathsf{A}}[U|V^{(n)}])_{n\in\mathbb{N}}$ *is a martingale with respect to the filtration* $(\sigma(V^{(n)}))_{n\in\mathbb{N}}$; *more precisely,*

$$\mathbb{E}\big[\mathbb{E}^{\mathsf{A}}[U|V] \mid V^{(n)}\big] = \mathbb{E}^{\mathsf{A}}\big[U|V^{(n)}\big] \quad a.s.; \tag{4.4}$$

(b) *the following martingale convergence theorem holds a.s. and in* $L^p(\mathbb{P}; \mathcal{G})$ *for* $1 \le p \le 2$:

$$\mathbb{E}^{\mathsf{A}}\big[U|V^{(n)}\big] \xrightarrow[n\to\infty]{} \mathbb{E}^{\mathsf{A}}[U|V]. \tag{4.5}$$

Theorem 4.13 (Formula for the LCE: incompatible case). *With the notation of Section 3 and using Notation 4.10, the operators* $\gamma_{U|V}^{(n)}$ *minimize the functional* $\mathcal{E}_{U|V}$ *given by (4.1) within* $\mathsf{A}(\mathcal{H}; \mathcal{G})$ *for* $n \to \infty$, *i.e.*

$$\mathcal{E}_{U|V}\big(\gamma_{U|V}^{(n)}\big) \xrightarrow[n\to\infty]{} \inf_{\gamma\in\mathsf{A}(\mathcal{H};\mathcal{G})} \mathcal{E}_{U|V}(\gamma).$$

In other words,

$$\big\|\mathbb{E}^{\mathsf{A}}[U|V] - \gamma_{U|V}^{(n)} \circ V\big\|_{L^2(\Omega,\Sigma,\mathbb{P};\mathcal{G})} \xrightarrow[n\to\infty]{} 0. \tag{4.6}$$

Further, denoting the pushforward of $\mathbb{P}$ *under* V *by* $\mathbb{P}_V$, $\gamma_{U|V}^{(n)} \circ V$ *converges to* $\mathbb{E}^{\mathsf{A}}[U|V]$ *a.s.,*

$$\big\|\gamma_{U|V}^{(n)}(v) - \mathbb{E}^{\mathsf{A}}[U|V = v]\big\|_{\mathcal{G}} \xrightarrow[n\to\infty]{} 0 \quad \text{for } \mathbb{P}_V\text{-a.e. } v \in \mathcal{H}. \tag{4.7}$$

4.4. Explicit formula for the LCE: Regularised case

In most practical applications the means and (cross-)covariance operators of U and V are not accessible explicitly, but have to be approximated empirically from data (in the simplest case, from independent and identically distributed samples $(u_n, v_n) \sim \mathbb{P}_{UV}$, where $\mathbb{P}_{UV}$ denotes the joint distribution of U and

V). Since the Moore–Penrose pseudo-inverse $C_V^\dagger$ shows an unstable behaviour when approximated empirically [22], Section SM2, it is typically replaced by its regularised version $(C_V + \varepsilon \operatorname{Id}_{\mathcal{H}})^{-1}$, where $\varepsilon > 0$ is a regularisation parameter. The following theorem shows that this is a principled way to address this issue, since the resulting operators minimize a perturbed functional $\mathcal{E}^{\mathrm{reg}}_{U|V}$. The natural space of operators in this context turns out to be the space of affine Hilbert–Schmidt operators.

Theorem 4.14 (Formula for the LCE: regularised case). *With the notation of Section 3, the operator* $\gamma_\varepsilon^{\mathsf{A}_2} \in \mathsf{A}_2(\mathcal{H}; \mathcal{G})$ *defined by*

$$\gamma_\varepsilon^{\mathsf{A}_2}(v) := \mu_U + C_{UV}(C_V + \varepsilon \operatorname{Id}_{\mathcal{H}})^{-1}(v - \mu_V)$$

minimizes the Tikhonov–Philipps-regularised functional

$$\mathcal{E}^{\mathrm{reg}}_{U|V}(\gamma) = \mathcal{E}_{U|V}(\gamma) + \varepsilon \|\gamma\|^2_{\mathsf{A}_2}$$

within $\mathsf{A}_2(\mathcal{H}; \mathcal{G})$, *where* $\mathcal{E}_{U|V}$ *is given by* (4.1).

4.5. Explicit formula for the linear conditional covariance

Before we derive a formula for the LCC $\mathbb{C}\mathrm{ov}^{\mathsf{A}}[U, W|V]$, the following theorem states an explicit formula for and some basic properties of the ALCC $\mathbb{C}\mathrm{ov}^{\mathsf{A}}_V[U, W]$. In particular, as mentioned earlier in Remark 4.2, we characterise the ALCC as the expected value $\mathbb{E}[\mathbb{C}\mathrm{ov}^{\mathsf{A}}[U, W|V]]$ of the LCC – hence our terminology for each of these conditional covariances.

Theorem 4.15 (Properties of the ALCC). *With the notation of Section 3 and using Notation 4.10,*

(a) $\mathbb{E}[\mathbb{C}\mathrm{ov}^{\mathsf{A}}[U, W|V]] = \mathbb{C}\mathrm{ov}^{\mathsf{A}}_V[U, W]$;
(b) $\mathbb{C}\mathrm{ov}^{\mathsf{A}}_V[U] = \mathbb{E}[\mathbb{C}\mathrm{ov}^{\mathsf{A}}[U|V]] \geq \mathbb{E}[\mathbb{C}\mathrm{ov}[U|V]] \geq 0$;
(c) $\mathbb{C}\mathrm{ov}^{\mathsf{A}}_V[U, W] = C_{UW} - M^*_{VU} M_{VW}$,
in particular, in the compatible case $\operatorname{ran} C_{VW} \subseteq \operatorname{ran} C_V$,
$\mathbb{C}\mathrm{ov}^{\mathsf{A}}_V[U, W] = C_{UW} - C_{UV} C_V^\dagger C_{VW}$.

Remark 4.16. The inequality $\mathbb{C}\mathrm{ov}^{\mathsf{A}}_V[U] \geq 0$ can also be seen more directly using (c):

$$\mathbb{C}\mathrm{ov}^{\mathsf{A}}_V[U] = C_U - M^*_{VU} M_{VU} = C_U - \left(R_{VU} C_U^{1/2}\right)^* R_{VU} C_U^{1/2} = C_U^{1/2}\left(\operatorname{Id}_{\mathcal{G}} - R^*_{VU} R_{VU}\right) C_U^{1/2} \geq 0,$$

where we used Notation 4.10 and $\|R_{VU}\| \leq 1$ (cf. Theorem A.3). While its expected value is always non-negative, the LCC $\mathbb{C}\mathrm{ov}^{\mathsf{A}}[U|V]$ itself can take on negative values, see Theorem 4.7(i).

Remark 4.17. Computing the conditional covariance $\mathbb{C}\mathrm{ov}[U|V = v]$ for various $v \in \mathcal{H}$ is often too costly, in which case one can focus on its mean $\mathbb{E}[\mathbb{C}\mathrm{ov}[U|V]]$. The above statements show that the average LCC $\mathbb{E}[\mathbb{C}\mathrm{ov}^{\mathsf{A}}[U|V]] = \mathbb{C}\mathrm{ov}^{\mathsf{A}}_V[U]$, which can be computed by the Gaussian conditional covariance formula, never underestimates the true expected conditional covariance $\mathbb{E}[\mathbb{C}\mathrm{ov}[U|V]]$.

As a consequence, we obtain an explicit formula for the LCC $\mathbb{C}\mathrm{ov}^{\mathsf{A}}[U, W|V]$:

Corollary 4.18 (Formula for the LCC). *With the notation of Section 3, let*

$$Z := R^{\mathsf{A}}[U|V] \otimes R^{\mathsf{A}}[W|V] \colon \Omega \to \mathsf{L}_2(\mathcal{F}; \mathcal{G}).$$

Then, using Notation 4.10,

$$\mathbb{C}\mathrm{ov}^{\mathsf{A}}[U, W|V] = C_{UW} - M_{VU}^{*} M_{VW} + \lim_{n\to\infty} \big(C_V^{(n)\dagger} C_{VZ}^{(n)}\big)^{*} (V - \mu_V) \quad a.s.,$$

where the limit is in the Hilbert–Schmidt norm and

$$\begin{aligned} \mu_Z &= C_{UW} - M_{VU}^{*} M_{VW}, \\ C_{VZ} &= \mathbb{E}\big[V \otimes \big(U - \mathbb{E}^{\mathsf{A}}[U|V]\big) \otimes \big(W - \mathbb{E}^{\mathsf{A}}[W|V]\big)\big] - \mu_V \otimes \mu_Z. \end{aligned}$$

In particular, in the compatible case with $\operatorname{ran} C_{VW} \subseteq \operatorname{ran} C_V$ *and* $\operatorname{ran} C_{VZ} \subseteq \operatorname{ran} C_V$,

$$\mathbb{C}\mathrm{ov}^{\mathsf{A}}[U, W|V] = C_{UW} - C_{UV} C_V^{\dagger} C_{VW} + \big(C_V^{\dagger} C_{VZ}\big)^{*} (V - \mu_V) \quad a.s.$$

5. Application to kernel conditional mean embeddings

The above results have a beautiful application to the derivation of conditional mean embeddings (CMEs), a concept used in machine learning to perform conditioning of random variables after embedding them into suitable reproducing kernel Hilbert spaces (RKHSs). To this end, let $\mathcal{H}$ and $\mathcal{G}$ be two RKHSs over measurable spaces $\mathcal{X}$ and $\mathcal{Y}$ respectively, with reproducing kernels k and ℓ and canonical feature maps $\varphi(x) := k(x, \bullet)$ and $\psi(y) := \ell(y, \bullet)$.

For two random variables $X\colon \Omega \to \mathcal{X}$ and $Y\colon \Omega \to \mathcal{Y}$ with joint distribution $\mathbb{P}_{XY}$ and corresponding marginal distributions $\mathbb{P}_X$ and $\mathbb{P}_Y$ such that $V := \varphi(X) \in L^2(\Omega, \Sigma, \mathbb{P}; \mathcal{H})$ and $U := \psi(Y) \in L^2(\Omega, \Sigma, \mathbb{P}; \mathcal{G})$, respectively, the CME $\mathbb{E}[U|X]$ can be characterised by the linear-algebraic transformation

$$\mathbb{E}[U|X] = \mu_U + \big(C_V^{\dagger} C_{VU}\big)^{*} \big(\varphi(X) - \mu_V\big), \tag{5.1}$$

which holds under appropriate technical assumptions [22]. Formula (5.1) can be interpreted as saying that application of the Kálmán update or BLUE formulae to RKHS embeddings of random variables realises the embedding of conditional distributions.

The theory on (affine) linear conditional means from Section 4 provides an alternative proof and a more insightful and explanatory derivation of (5.1). The main idea is to find conditions under which the conditional expectation $\mathbb{E}[U|V]$ agrees with the linear conditional expectation $\mathbb{E}^{\mathsf{A}}[U|V]$. Assuming φ to be injective implies that $\mathbb{E}[U|X] = \mathbb{E}[U|V]$ and (5.1) then follows directly from Theorem 4.8, while Theorem 4.13 implies the more generally applicable formula in Theorem 5.11. The reason why one could hope for $\mathbb{E}[U|V] = \mathbb{E}^{\mathsf{A}}[U|V]$ to hold is the celebrated *kernel trick*, the guiding theme of RKHS-based methods: many nonlinear problems in the original spaces $\mathcal{X}$ and $\mathcal{Y}$ (here, conditioning) become linear-algebraic problems when embedded into the corresponding RKHSs $\mathcal{H}$ and $\mathcal{G}$.

5.1. Setup and notation

Here, with apologies for the large notational overhead relative to the brevity of the results in Section 5.2, we introduce the precise technical assumptions and notation needed for the validity of the CME approach; see [22], Section 2, for a detailed exposition.

Regarding the kernel mean embedding of random variables $X\colon \Omega \to \mathcal{X}$ and $Y\colon \Omega \to \mathcal{Y}$ into RKHSs $\mathcal{H}$ and $\mathcal{G}$ over $\mathcal{X}$ and $\mathcal{Y}$, respectively, with reproducing kernels k and ℓ via the canonical feature maps $\varphi(x) := k(x, \bullet)$ and $\psi(y) := \ell(y, \bullet)$ we make the following basic assumptions:

Assumption 5.1.

(a) The space $\mathcal{X}$ is a measurable space and $\mathcal{Y}$ is a Borel space.
(b) The kernel functions $k\colon \mathcal{X}\times\mathcal{X}\to\mathbb{R}$ and $\ell\colon \mathcal{Y}\times\mathcal{Y}\to\mathbb{R}$ are symmetric positive definite and measurable and the corresponding RKHSs $(\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}})$ and $(\mathcal{G}, \langle\cdot,\cdot\rangle_{\mathcal{G}})$ are separable. Moreover, the canonical feature map $\varphi(x) := k(x,\cdot)$ is injective.[6]
(c) The random variables $V := \varphi(X)$ and $U := \psi(Y)$ lie in $L^2(\Omega,\Sigma,\mathbb{P};\mathcal{G})$ and $L^2(\Omega,\Sigma,\mathbb{P};\mathcal{H})$, respectively.
(d) For any $h\in\mathcal{H}$ we have $\|h\|_{\mathcal{H}} = 0$ if and only if $h = 0$ $\mathbb{P}_X$-a.e. in $\mathcal{X}$.

By [21], Theorem 5.3, Assumption 5.1(a) ensures the existence of a $\mathbb{P}_X$-a.e.-unique regular version of the conditional probability distribution $\mathbb{P}_{Y|X=x}$, $x\in\mathcal{X}$, for random variables $X\colon\Omega\to\mathcal{X}$ and $Y\colon\Omega\to\mathcal{Y}$. Moreover, by [35], Lemma 4.25, Assumption 5.1(b) guarantees the measurability of $\varphi\colon\mathcal{X}\to\mathcal{H}$ and $\psi\colon\mathcal{Y}\to\mathcal{G}$, respectively. Assumption 5.1(c) implies that $\mathcal{H}$ (resp. $\mathcal{G}$) is continuously embedded in the pre-Hilbert space $\mathcal{L}^2(\mathbb{P}_X)$ (resp. $\mathcal{L}^2(\mathbb{P}_Y)$). Furthermore, it also follows that $\mathbb{E}[\|\psi(Y)\|_{\mathcal{G}}^2|X=x]<\infty$ and that $\mathcal{G}$ is continuously embedded in $\mathcal{L}^2(\mathbb{P}_{Y|X=x})$ for all $x\in\mathcal{X}_Y$, where $\mathcal{X}_Y\subseteq\mathcal{X}$ has full $\mathbb{P}_X$ measure, see [22], Section 2. Assumption 5.1(d) clearly holds if k is continuous and $\operatorname{supp}(\mathbb{P}_X)=\mathcal{X}$. In particular, it allows us to view $\mathcal{H}$ as a subspace of the Lebesgue Hilbert space $L^2(\mathbb{P}_X)$.

Subsequently, we work with Bochner spaces $L^2(\mathbb{P}_X;\mathcal{F})$ where $\mathcal{F}$ denotes another separable Hilbert space (which in our case will be equal to either $\mathbb{R}$ or $\mathcal{G}$). Recall that the space $L^2(\mathbb{P}_X;\mathcal{F})$ is isometrically isomorphic to the Hilbert tensor product space $\mathcal{F}\otimes L^2(\mathbb{P}_X)$. We comment on the various perspectives on tensor product space exploited in our proofs in Remark 5.3 below. For stating our second main result, we require the following definitions.

Notation 5.2.

(a) Given a separable Hilbert space $\mathcal{F}$ we define $L^2_{\mathcal{C}}(\mathbb{P}_X;\mathcal{F})$ to be the quotient space $L^2(\mathbb{P}_X;\mathcal{F})/\mathcal{C}$, where

$$\mathcal{C} := \big\{ f\in L^2(\mathbb{P}_X;\mathcal{F}) \mid \exists c\in\mathcal{F}\colon\ f(x)=c \text{ for } \mathbb{P}_X\text{-a.e. } x\in\mathcal{X}\big\},$$
$$\big\langle [f_1],[f_2]\big\rangle_{L^2_{\mathcal{C}}(\mathbb{P}_X;\mathcal{F})} := \big\langle f_1-\mathbb{E}\big[f_1(X)\big],\, f_2-\mathbb{E}\big[f_2(X)\big]\big\rangle_{L^2(\mathbb{P}_X;\mathcal{F})}.$$

In the case $\mathcal{F}=\mathbb{R}$, we abbreviate the space $L^2_{\mathcal{C}}(\mathbb{P}_X;\mathbb{R})$ by $L^2_{\mathcal{C}}(\mathbb{P}_X)$ and for any subspace $\mathcal{U}\subseteq L^2(\mathbb{P}_X;\mathcal{F})$ we define $\mathcal{U}_{\mathcal{C}} := \mathcal{U}/(\mathcal{U}\cap\mathcal{C})$ and identify it with a subspace of $L^2_{\mathcal{C}}(\mathbb{P}_X;\mathcal{F})$.

(b) Furthermore, we define the main object of our interest, the conditional mean

$$\mathfrak{m}\colon\mathcal{X}\to\mathcal{G},\qquad \mathfrak{m}(x) := \begin{cases}\mathbb{E}[U|X=x], & \text{for } x\in\mathcal{X}_Y,\\ 0, & \text{otherwise.}\end{cases}$$

Note that $\mathfrak{m}\in L^2(\mathbb{P}_X;\mathcal{G})$, since Jensen's inequality, the law of total expectation, and Assumption 5.1(c) together yield that

$$\|\mathfrak{m}\|^2_{L^2(\mathbb{P}_X;\mathcal{G})} = \mathbb{E}\big[\big\|\mathbb{E}\big[\psi(Y)|X\big]\big\|_{\mathcal{G}}^2\big] \le \mathbb{E}\big[\mathbb{E}\big[\|U\|_{\mathcal{G}}^2|X\big]\big] = \mathbb{E}\big[\|U\|_{\mathcal{G}}^2\big] < \infty.$$

[6] The injectivity of φ was not required for the derivations in [22]. However, it represents a minor restriction since one typically considers characteristic kernels, which implies injectivity of φ.

(c) We also introduce the notation

$$f_g(x) := \mathbb{E}\big[g(Y)|X = x\big] = \big\langle g, \mathfrak{m}(x)\big\rangle_{\mathcal{G}},$$

mainly for the comparison of our formulations to [22].

Remark 5.3. For $\mathcal{F} = \mathcal{H}$ and $\mathcal{F} = L^2(\mathbb{P}_X)$, we will view the Hilbert tensor product space $\mathcal{G} \otimes \mathcal{F}$ from three[7] perspectives:

- $\mathcal{G} \otimes \mathcal{F}$ is isometrically isomorphic to the space $\mathsf{L}_2(\mathcal{F}; \mathcal{G})$ of Hilbert–Schmidt operators from $\mathcal{F}$ to $\mathcal{G}$ [2], Chapter 12, and we sometimes view $g \otimes f \in \mathcal{G} \otimes \mathcal{F} \cong \mathsf{L}_2(\mathcal{F}; \mathcal{G})$ as the corresponding mapping from $\mathcal{F}$ to $\mathcal{G}$ given by

$$[g \otimes f]_{\mathcal{F}\to\mathcal{G}}\big(f'\big) := \big\langle f, f'\big\rangle_{\mathcal{F}} g.$$

- $\mathcal{G} \otimes \mathcal{F}$ can be viewed as a set of functions from $\mathcal{X}$ to $\mathcal{G}$ [2], Chapter 12. Thus, we can view $g \otimes f \in \mathcal{G} \otimes \mathcal{F} \subseteq L^2(\mathbb{P}_X; \mathcal{G})$ accordingly as a function on $\mathcal{X}$ taking values in $\mathcal{G}$:

$$[g \otimes f]_{\mathcal{X}\to\mathcal{G}}(x) := f(x)g = [g \otimes f]_{\mathcal{H}\to\mathcal{G}}\big(\varphi(x)\big),$$

where the last equality holds for $\mathcal{F} = \mathcal{H}$ by the reproducing property, in which case we have established the important observation

$$\mathfrak{f}_{\mathcal{X}\to\mathcal{G}}(x) = \mathfrak{f}_{\mathcal{H}\to\mathcal{G}}\big(\varphi(x)\big), \quad \mathfrak{f} \in \mathcal{G} \otimes \mathcal{H}, x \in \mathcal{X}. \tag{5.2}$$

Note that $\mathcal{G} \otimes \mathcal{F}$ is indeed a subspace of $L^2(\mathbb{P}_X; \mathcal{G})$, which is obvious in the case $\mathcal{F} = L^2(\mathbb{P}_X)$, and, in the case $\mathcal{F} = \mathcal{H}$, follows from

$$\|\mathfrak{f}_{\mathcal{X}\to\mathcal{G}}\|^2_{L^2(\mathbb{P}_X;\mathcal{G})} = \mathbb{E}\big[\big\|\mathfrak{f}_{\mathcal{X}\to\mathcal{G}}(X)\big\|^2_{\mathcal{G}}\big] \overset{(5.2)}{\leq} \|\mathfrak{f}_{\mathcal{H}\to\mathcal{G}}\|^2_{\mathsf{L}_2}\mathbb{E}\big[\big\|\varphi(X)\big\|^2_{\mathcal{H}}\big] < \infty,$$

where we used Assumption 5.1(c) in the last step. This is hardly surprising, since $\mathcal{H} \subseteq L^2(\mathbb{P}_X)$ by Assumption 5.1(c), (d), but not trivial as we use the RKHS norm $\|\bullet\|_{\mathcal{H}}$ in the construction of $\mathcal{G} \otimes \mathcal{H}$, which may not agree with $\|\bullet\|_{L^2(\mathbb{P}_X)}$.
- Since tensor products are commutative up to isometric isomorphism, $\mathcal{G} \otimes \mathcal{F}$ is also isometrically isomorphic to $\mathsf{L}_2(\mathcal{G}; \mathcal{F})$ and we can analogously set

$$[g \otimes f]_{\mathcal{G}\to\mathcal{F}}\big(g'\big) := \big\langle g, g'\big\rangle_{\mathcal{F}} f, \quad f \in \mathcal{F}, g, g' \in \mathcal{G}.$$

We will sometimes use the resulting identities for arbitrary $\mathfrak{f} \in \mathcal{G} \otimes \mathcal{F}$: with $x \in \mathcal{X}$

$$\mathfrak{f}_{\mathcal{G}\to\mathcal{F}}(g)(x) = \big\langle \mathfrak{f}_{\mathcal{X}\to\mathcal{G}}(x), g\big\rangle_{\mathcal{G}}, \qquad \langle \mathfrak{f}, g \otimes f\rangle_{\mathcal{G}\otimes\mathcal{F}} = \big\langle \mathfrak{f}_{\mathcal{G}\to\mathcal{F}}(g), f\big\rangle_{\mathcal{F}}. \tag{5.3}$$

However, we will drop the indices $\mathcal{F} \to \mathcal{G}$, $\mathcal{X} \to \mathcal{G}$ and $\mathcal{G} \to \mathcal{F}$ in the following, since it will always be clear which version we mean, whenever we apply $\mathfrak{f} \in \mathcal{G} \otimes \mathcal{F}$ to some element of $\mathcal{F}$, $\mathcal{X}$, or $\mathcal{G}$, respectively.

[7] In fact, there is another viewpoint on $\mathcal{G} \otimes \mathcal{F}$, namely as a set of functions from $Y \times X$ to $\mathbb{R}$, where $(g \otimes f)(y, x) := g(y) f(x)$, in which case $\mathcal{G} \otimes \mathcal{F} \subseteq L^2(\mathbb{P}_Y \otimes \mathbb{P}_X)$.

The typical assumption for CMEs is that the functions f_g introduced in Notation 5.2(c) must lie in $\mathcal{H}$ for all $g \in \mathcal{G}$. [22] discuss several weaker assumptions on f_g, which we are going to adopt in this paper. However, the main purpose of using f_g is that, for $g = \psi(y)$ with $y \in \mathcal{Y}$, $f_{\psi(y)} = \mu_{Y|X=\cdot}(y)$ and, in fact, all results in [22] rely solely on these special cases of f_g. It is therefore meaningful to restate these assumptions in terms of $\mathfrak{m}$ rather than f_g.

Assumption 5.4. Under Assumption 5.1 and using Notation 5.2, we introduce the following assumptions on the functions $\mathfrak{m} \in L^2(\mathbb{P}_X; \mathcal{G})$:

(A) $\mathfrak{m} \in \mathcal{G} \otimes \mathcal{H}$;
(B) $[\mathfrak{m}] \in (\mathcal{G} \otimes \mathcal{H})_{\mathcal{C}}$;
(C) $P_{\overline{(\mathcal{G}\otimes\mathcal{H})_{\mathcal{C}}}^{L^2_{\mathcal{C}}(\mathbb{P}_X;\mathcal{G})}}[\mathfrak{m}] \in (\mathcal{G} \otimes \mathcal{H})_{\mathcal{C}}$;
(^{u}C) $P_{\overline{\mathcal{G}\otimes\mathcal{H}}^{L^2(\mathbb{P}_X;\mathcal{G})}}\mathfrak{m} \in \mathcal{G} \otimes \mathcal{H}$;
(A*) $\mathfrak{m} \in \overline{\mathcal{G} \otimes \mathcal{H}}^{L^2(\mathbb{P}_X;\mathcal{G})}$;
(B*) $[\mathfrak{m}] \in \overline{(\mathcal{G} \otimes \mathcal{H})_{\mathcal{C}}}^{L^2_{\mathcal{C}}(\mathbb{P}_X;\mathcal{G})}$.

Remark 5.5. Note that these assumptions are slightly stronger than the corresponding assumptions on f_g in [22], Section 3. The corresponding implications are formulated in Proposition 5.6 below. However, by Lemma 5.7, (B*) still follows from k being characteristic, providing a verifiable condition for the applicability of the corresponding CME formula (Theorem 5.11). Also, as before, (A*) follows from k being L^2-universal. Indeed, if $\mathcal{G} \otimes \mathcal{H}$ is dense in $L^2(\mathbb{P}_X; \mathcal{G})$, then

$$\overline{\mathcal{G} \otimes \mathcal{H}}^{L^2(\mathbb{P}_X;\mathcal{G})} \subseteq \overline{\mathcal{G} \otimes \overline{\mathcal{H}}^{L^2(\mathbb{P}_X)}}^{\mathcal{G}\otimes L^2(\mathbb{P}_X)} = \overline{\mathcal{G} \otimes L^2(\mathbb{P}_X)}^{\mathcal{G}\otimes L^2(\mathbb{P}_X)} = L^2(\mathbb{P}_X; \mathcal{G}).$$

These and further relations among the conditions in Assumption 5.4 are summarised in Figure 3.

Proposition 5.6. *The conditions in Assumption 5.4 imply the corresponding assumptions on f_g in* [22], *Section* 3, (*here marked with a subscript "old"*). *More precisely, under Assumption 5.1 and using Notation 5.2,*

(A) $\implies$ (A$_{\text{old}}$) $f_g \in \mathcal{H}$ *for each* $g \in \mathcal{G}$;
(B) $\implies$ (B$_{\text{old}}$) $[f_g] \in \mathcal{H}_{\mathcal{C}}$ *for each* $g \in \mathcal{G}$;
(C) $\implies$ (C$_{\text{old}}$) $P_{\overline{\mathcal{H}_{\mathcal{C}}}^{L^2_{\mathcal{C}}(\mathbb{P}_X)}}[f_g] \in \mathcal{H}_{\mathcal{C}}$ *for each* $g \in \mathcal{G}$;
(^{u}C) $\implies$ (^{u}C$_{\text{old}}$) $P_{\overline{\mathcal{H}}^{L^2(\mathbb{P}_X)}} f_g \in \mathcal{H}$ *for each* $g \in \mathcal{G}$;
(A*) $\implies$ (A$^*_{\text{old}}$) $f_g \in \overline{\mathcal{H}}^{L^2(\mathbb{P}_X)}$ *for each* $g \in \mathcal{G}$;
(B*) $\implies$ (B$^*_{\text{old}}$) $f_g \in \overline{(\mathcal{H})_{\mathcal{C}}}^{L^2_{\mathcal{C}}(\mathbb{P}_X)}$ *for each* $g \in \mathcal{G}$.

Further,

(C) $\implies$ $\operatorname{ran} C_{VU} \subseteq \operatorname{ran} C_V$;
(^{u}C) $\implies$ $\operatorname{ran}{}^u C_{VU} \subseteq \operatorname{ran}{}^u C_V$.

Lemma 5.7. *Under Assumption 5.1 and using Notation 5.2, if k is a characteristic kernel, then $(\mathcal{G} \otimes \mathcal{H})_{\mathcal{C}}$ is dense in $L^2_{\mathcal{C}}(\mathbb{P}_X; \mathcal{G})$ and Assumption 5.4(B*) is satisfied.*

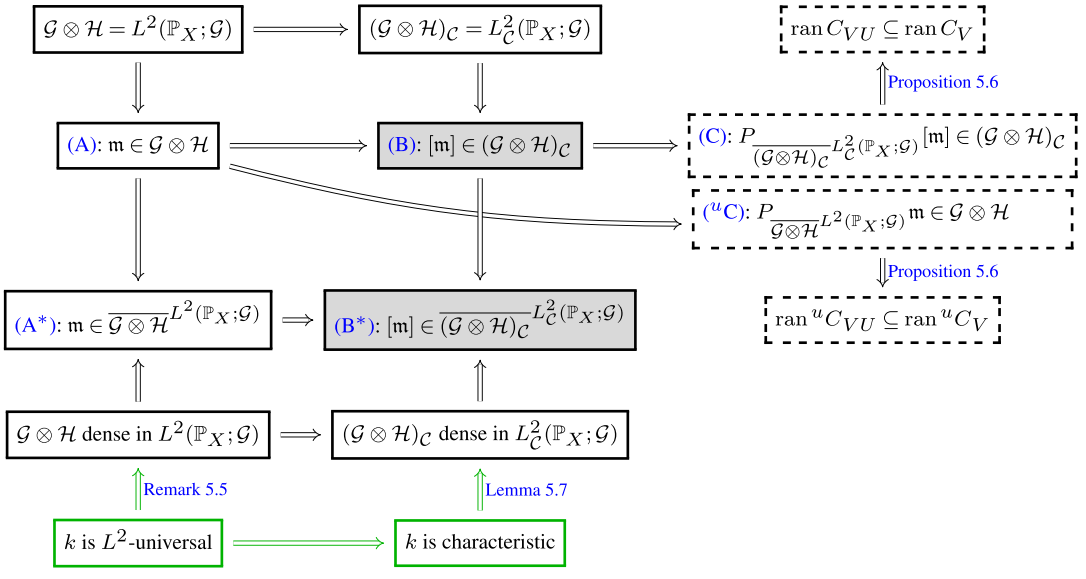

Figure 3. A hierarchy of CME-related assumptions. Sufficient conditions for validity of the CME formula are indicated by solid boxes while the insufficient Assumptions (C) and (^{u}C), indicated by dashed boxes, have several strong theoretical implications. Assumption 5.4(B*) is the most favorable one, since it is verifiable in practice, and, by Lemma 5.7, in particular is fulfilled if the kernel is universal or even just characteristic (marked in green). The shaded boxes correspond to Theorems 5.8 and 5.11.

5.2. Derivation of the CME formula

We are now in a position to rederive the CME formula under Assumption 5.4.

Theorem 5.8 (CME under Assumption 5.4(A) or (B)). *Under Assumptions 5.1 and 5.4(B) the operator $C_V^\dagger C_{VU}\colon\mathcal{G}\to\mathcal{H}$ is bounded and*

$$\mathbb{E}[U|X]=\mu_U+\big(C_V^\dagger C_{VU}\big)^*\big(\varphi(X)-\mu_V\big)\quad a.s. \tag{5.4}$$

Remark 5.9. Note that we have proven a stronger statement than originally intended. Namely, the CME operator $(C_V^\dagger C_{VU})^*$ is not just bounded, but even Hilbert–Schmidt. However, it can be argued that this property is already hidden in the assumptions, namely in Assumption 5.4(B), since $\mathcal{G}\otimes\mathcal{H}\cong\mathsf{L}_2(\mathcal{H};\mathcal{G})$;

Remark 5.10. A similar statement can be proven for uncentered covariance operators under the stronger Assumption 5.4(A), see [22], Theorem 5.3.

Theorem 5.11 (CME under Assumption 5.4(B*)). *Under Assumptions 5.1 and 5.4(B*) and using Notation 4.10, the operators $\gamma_{U|V}^{(n)}$ satisfy*

$$\big\|\mathbb{E}[U|X]-\gamma_{U|V}^{(n)}\big(\varphi(X)\big)\big\|_{L^2(\mathbb{P};\mathcal{G})}\xrightarrow[n\to\infty]{}0,$$

$$\big\|\mathbb{E}[U|X=x]-\gamma_{U|V}^{(n)}\big(\varphi(x)\big)\big\|_{\mathcal{G}}\xrightarrow[n\to\infty]{}0\quad \textit{for } \mathbb{P}_X\textit{-a.e. } x\in\mathcal{X}.$$

6. Application to Gaussian conditioning in Hilbert spaces

While the conditioning of a Gaussian random variable $(U, V)\colon \Omega \to \mathcal{G} \oplus \mathcal{H}$ on its second component is a well-established concept [19,27], the most general case (where C_V is not necessarily injective) has only been treated rather recently by [29]. In that work, by developing an approximation theory for shorted operators in terms of oblique projections and applying the martingale convergence theorem, the authors derive approximating sequences for both the conditional expectation $\mathbb{E}[U|V]$ and the conditional covariance operator $\mathbb{C}\mathrm{ov}[U|V]$.

The formula that [29], Theorem 3.3, obtain for the conditional expectation $\mathbb{E}[U|V]$ is identical to (4.7) (with $\mathbb{E}[U|V]$ in place of $\mathbb{E}^{\mathsf{A}}[U|V]$). Similar to CMEs in Section 5, our theory provides an alternative derivation of this formula by

(i) proving $\mathbb{E}[U|V] \in \overline{\mathsf{A}_{\mathcal{G}} \circ V}$, implying the identity $\mathbb{E}[U|V] = \mathbb{E}^{\mathsf{A}}[U|V]$;
(ii) applying Theorem 4.13.

Let us give a short sketch of the proof:

Proof sketch for (i). Let $\overline{U} := U - \mu_U$, $\overline{V} := V - \mu_V$ and $\gamma := \gamma_{\overline{U}|\overline{V}}\colon \mathcal{H} \to \mathcal{G}$ be the corresponding CEF. [36], Theorem 3.11, show that there exists a linear subspace $\tilde{\mathcal{H}}$ of $\mathcal{H}$ such that $\overline{V} \in \tilde{\mathcal{H}}$ a.s. and the restriction $\gamma|_{\tilde{\mathcal{H}}}$ is linear. In the proof, the authors further construct a sequence $\gamma_n \in \mathsf{L}(\mathcal{H}; \mathcal{G})$ such that

$$\gamma_n(h) = \sum_{i=1}^{n} \langle h, h_i \rangle \gamma(h_i) \xrightarrow[n\to\infty]{} \gamma(h) \quad \text{for all } h \in \tilde{\mathcal{H}}.$$

Using the Karhunen–Loève expansion of $\overline{V}$, one can prove

$$\|\gamma_n \circ \overline{V} - \gamma \circ \overline{V}\|^2_{L^2(\mathbb{P};\mathcal{G})} \xrightarrow[n\to\infty]{} 0.$$

Hence $\mathbb{E}[\overline{U}|\overline{V}] = \gamma \circ \overline{V} \in \overline{\mathsf{L}_{\mathcal{G}} \circ V}$ and thereby $\mathbb{E}[U|V] = \mu_U + \gamma \circ (V - \mu_V) \in \overline{\mathsf{A}_{\mathcal{G}} \circ V}$. □

The appeal to [36], Theorem 3.11, is somewhat unsatisfactory, since close inspection of the proof of that theorem reveals that it in fact establishes the entire conditional mean formula for Gaussian conditioning. In this sense, our derivation of the formula for the conditional *mean* $\mathbb{E}[U|V]$ in the Gaussian case is not novel. However, let us now turn to the conditional *covariance* $\mathbb{C}\mathrm{ov}[U|V]$.

It is well known that the conditional covariance is constant for Gaussian random variables (i.e. it does not depend on the value of the conditioning variable), and that, since $\mathbb{E}[U|V] = \mathbb{E}^{\mathsf{A}}[U|V]$, it coincides with the ALCC, $\mathbb{C}\mathrm{ov}[U|V] = \mathbb{C}\mathrm{ov}^{\mathsf{A}}_V[U]$. Therefore, our results show that, in contrast to the conditional expectation, the conditional covariance does require the approximating sequence established by [29], Theorem 3.4, but the explicit formula from Theorem 4.15(c) applies. In summary, we can consider three versions of the Gaussian conditional covariance formula:

- the *invertible case* in which C_V is invertible and, in particular, $\mathcal{H}$ is finite dimensional:

$$\mathbb{C}\mathrm{ov}[U|V] = C_U - C_{UV} C_V^{-1} C_{VU}.$$

- the *compatible case* in which $\operatorname{ran} C_{VU} \subseteq \operatorname{ran} C_V$:
  By Theorem A.1, the operator $C_V^{\dagger} C_{VU} \in \mathsf{L}(\mathcal{G}; \mathcal{H})$ is well defined and bounded and

$$\mathbb{C}\mathrm{ov}[U|V] = C_U - C_{UV} C_V^{\dagger} C_{VU}.$$

- the *incompatible* (or *general*) *case*:
  By Theorem A.3, the operator $M_{VU} := (C_V^{1/2})^\dagger C_{VU}$ is well-defined and bounded and

$$\mathbb{C}\mathrm{ov}[U|V] = C_U - M_{VU}^* M_{VU}.$$

7. Closing remarks

This paper presents a rigorous theory of the linear conditional expectation (LCE) $\mathbb{E}^{\mathsf{A}}[\bullet|V]$ that strongly extends the existing theory on Bayes linear analysis [18].

After the definitions of the linear conditional expectation $\mathbb{E}^{\mathsf{A}}[U|V]$ and the linear conditional covariance (LCC) operator $\mathbb{C}\mathrm{ov}^{\mathsf{A}}[U, W|V]$ – which is related to, but differs from, the so-called adjusted covariance used in Bayes linear statistics – we studied in detail which properties of the common conditional expectation $\mathbb{E}[\bullet|V]$ and conditional covariance $\mathbb{C}\mathrm{ov}[\bullet, \bullet|V]$ hold for their linear approximations. Amongst others, we proved several tower properties and the laws of total expectation and total covariance. On the other hand, $\mathbb{E}^{\mathsf{A}}[\bullet|V]$ is neither monotonic, nor contractive in $L^p(\mathbb{P}; \mathcal{G})$ (except for $p = 2$, which is clear from its definition) and does not fulfill the triangle inequality. The dominated convergence theorem holds only under modified assumptions and, so far, could only be proved under the assumption that C_V has finite rank (see Theorem 4.5(i), Remark 4.6, and Theorem 4.7(g)).

We derived explicit formulae for both the LCE and the LCC, distinguishing between the so-called compatible (simple) and the incompatible (hard) case, as well as providing a regularised formula for the LCE.

Naturally, whenever $\mathbb{E}^{\mathsf{A}}[U|V] = \mathbb{E}[U|V]$, these formulae apply for the common conditional expectation. This trivial observation allowed us to provide an alternative derivation of the Gaussian conditioning formulae and give a simple and intuitive proof of the widely-used technique of conditional mean embeddings (CMEs) in machine learning: it turns out that, if $U = \psi(Y)$ and $V = \varphi(X)$ are reproducing kernel Hilbert space (RKHS) embeddings of some random variables X and Y, then the above property holds true under rather mild conditions.

One direction for future work is the derivation of optimal regularisation schemes $\varepsilon(n) \to 0$ when the regularised case considered in Section 4.4 is applied to empirical sample data consisting of $n \to \infty$ data points. We anticipate that this will be a rich vein of research, but also decidedly non-trivial, since such problems admit no general solution and effective strategies (be they a priori, a posteriori, or heuristic) rely on appropriate source conditions for the unknowns.

Finally, we note that this work has concentrated on *centred* (cross-)covariance operators, which are associated with affine approximations of the conditional expectation function. Some, but not all of the statements can also be proved for *uncentered* operators, which are associated with linear approximations of the CEF and are often used in practice; for example, the uncentered formulation is commonly used for CMEs. However, the theory with uncentered operators has weaker statements and is more restrictive, and so we strongly encourage the use of centred operators.

8. Proofs

Proof of Proposition 3.2. We only give the proof of the second statement, which is similar to the first one but slightly more technical. Let $W \in \overline{\mathsf{A}_{\mathcal{G}} \circ V}$ and $(\gamma_n)_{n\in\mathbb{N}}$ be a sequence in $\mathsf{A}(\mathcal{H}; \mathcal{G})$, such that

$$\|\gamma_n \circ V - W\|_{L^2(\mathbb{P};\mathcal{G})} \xrightarrow[n\to\infty]{\text{a.s.}} 0.$$

This implies that, for $\overline{V} := V - \mu_V$ and $\overline{W} := W - \mu_W$,

$$\|\overline{\gamma}_n \circ \overline{V} - \overline{W}\|_{L^2(\mathbb{P};\mathcal{G})} \xrightarrow[n\to\infty]{\text{a.s.}} 0.$$

By [15], Corollary 2.32, there exists a subsequence $(\overline{\gamma}_{n_k})_{k\in\mathbb{N}}$ of $(\overline{\gamma}_n)_{n\in\mathbb{N}}$ such that

$$\big\|\overline{\gamma}_{n_k} \circ \overline{V}(\omega) - \overline{W}(\omega)\big\|_{\mathcal{G}} \xrightarrow[k\to\infty]{} 0 \quad \text{for } \mathbb{P}\text{-a.e. } \omega \in \Omega.$$

Define the linear subspace $\mathcal{H}_0 \subseteq \mathcal{H}$ and the (possibly unbounded) linear operator $A_0 \colon \mathcal{H}_0 \to \mathcal{G}$ by

$$\mathcal{H}_0 := \big\{h \in \mathcal{H} \mid \overline{\gamma}_{n_k}(h) \text{ converges in } \mathcal{G}\big\}, \qquad A_0(h) := \lim_{k\to\infty} \overline{\gamma}_{n_k}(h) \text{ for } h \in \mathcal{H}_0,$$

and extend A_0 trivially[8] to a linear operator A on $\mathcal{H}$. Then $\overline{V} \in \mathcal{H}_0$ a.s. and $A \circ \overline{V} = \overline{W}$ a.s. Considering the affine operator $\gamma \colon \mathcal{H} \to \mathcal{G}$ given by $\gamma(h) = \mu_W + A(h - \mu_V)$ yields that $\gamma \in \mathsf{A}_V(\mathcal{H};\mathcal{G})$ and $\gamma \circ V = W$ a.s. □

Proof of Lemma 4.4. Since $\mathbb{E}[\mathbb{E}^{\mathsf{A}}[U|V]] = \mathbb{E}[U]$ (which follows from $\mathbb{E}$ and $\mathbb{E}^{\mathsf{A}}[\bullet|V]$ being orthogonal projections, see the law of total linear expectation in Theorem 4.5(d)), it follows that $\mathbb{E}[R^{\mathsf{A}}[U|V]] = 0$. Hence, by Lemma A.6, for any $\gamma \in \mathsf{L}(\mathcal{H};\mathcal{G})$,

$$0 = \big\langle R^{\mathsf{A}}[U|V], \gamma \circ V\big\rangle_{L^2(\mathbb{P};\mathcal{G})} = \operatorname{tr}\big(\mathbb{C}\mathrm{ov}\big[R^{\mathsf{A}}[U|V], V\big]\gamma^*\big).$$

By Lemma A.5 this implies that $\mathbb{C}\mathrm{ov}[R^{\mathsf{A}}[U|V], V] = 0$. Now let $W \in \overline{\mathsf{A}_{\mathcal{F}} \circ V}$. By Proposition 3.2, $W = \gamma \circ V$ for some $\gamma \in \mathsf{A}_V(\mathcal{H};\mathcal{F})$. Hence, invoking Lemma A.6 another time,

$$\mathbb{C}\mathrm{ov}\big[R^{\mathsf{A}}[U|V], W\big] = \big(\overline{\gamma}\mathbb{C}\mathrm{ov}\big[V, R^{\mathsf{A}}[U|V]\big]\big)^* = 0,$$

which completes the proof. (Note that the finite trace of the cross-covariance operator was essential in the above argument.) □

Proof of Theorem 4.5. Properties (a)–(f), except for the second statement on linearity in (b) and the second tower property in (f), follow directly from the definitions of $\mathbb{E}[U]$, $\mathbb{E}[U|V]$ and $\mathbb{E}^{\mathsf{A}}[U|V]$ as orthogonal projections of U, the identity $\mathbb{E}[\langle \bullet, \bullet\rangle_{\mathcal{G}}] = \langle \bullet, \bullet\rangle_{L^2(\mathbb{P};\mathcal{G})}$ and the inclusions

$$\mathsf{A}(\mathcal{H};\mathcal{G}) \circ V \subseteq L^2\big(\sigma(V); G\big) \subseteq L^2(\Sigma; G), \qquad \mathsf{A}(\mathcal{F};\mathcal{G}) \circ \varphi \subseteq \mathsf{A}(\mathcal{H};\mathcal{G}).$$

For the second statement on linearity in (b) first note that $\psi(\mathbb{E}^{\mathsf{A}}[U|V]) \in \overline{\mathsf{A}_{\mathcal{F}} \circ V}$. Lemmas A.6 and 4.4 imply that, for any $\gamma \in \mathsf{A}(\mathcal{H};\mathcal{F})$,

$$\big\langle \psi(U) - \psi\big(\mathbb{E}^{\mathsf{A}}[U|V]\big), \gamma \circ V\big\rangle_{L^2(\mathbb{P};\mathcal{F})} = \operatorname{tr}\big(\psi\mathbb{C}\mathrm{ov}\big[\mathbb{E}^{\mathsf{A}}[U|V], V\big]\overline{\gamma}^* - \psi\mathbb{C}\mathrm{ov}[U, V]\overline{\gamma}^*\big) = 0,$$

which completes the proof of (b).

For second tower property in (f) let $\mathbb{E}^{\mathsf{A}}[U|V] = \gamma \circ V \in \overline{\mathsf{A}_{\mathcal{G}} \circ V}$ (using Proposition 3.2) and assume that there exists $\delta \in \mathsf{A}(\mathcal{G};\mathcal{G})$ such that

$$\big\|U - \delta \circ \mathbb{E}^{\mathsf{A}}[U|V]\big\|_{L^2(\mathbb{P};\mathcal{G})} < \big\|U - \mathbb{E}^{\mathsf{A}}[U|V]\big\|_{L^2(\mathbb{P};\mathcal{G})}.$$

[8] Choose a Hamel basis $\mathcal{B}_1$ of $\mathcal{H}_0$, extend it to a basis $\mathcal{B}_1 \cup \mathcal{B}_2$ of $\mathcal{H}$ and set $A := \tilde{A}$ on $\mathcal{B}_1$ and $A := 0$ on $\mathcal{B}_2$.

Then $\delta \circ \gamma \circ V \in \overline{\mathsf{A}_{\mathcal{G}} \circ V}$ is a better $L^2(\mathbb{P}; \mathcal{G})$-approximation of U than $\mathbb{E}^{\mathsf{A}}[U|V]$, which contradicts the definition of $\mathbb{E}^{\mathsf{A}}[U|V]$.

For the law of total linear covariance (g), first note that, by the law of total linear expectation (d) and Lemma 4.4,

$$\mathbb{E}\big[\mathbb{C}\mathrm{ov}^{\mathsf{A}}[U, W|V]\big] = \mathbb{E}\big[\mathbb{E}^{\mathsf{A}}\big[R^{\mathsf{A}}[U|V] \otimes R^{\mathsf{A}}[W|V] \mid V\big]\big] = \mathbb{C}\mathrm{ov}\big[R^{\mathsf{A}}[U|V], R^{\mathsf{A}}[W|V]\big].$$

Hence, again by Lemma 4.4,

$$\begin{aligned}\mathbb{C}\mathrm{ov}[U, W] &= \mathbb{C}\mathrm{ov}\big[\mathbb{E}^{\mathsf{A}}[U|V] + R^{\mathsf{A}}[U|V], \mathbb{E}^{\mathsf{A}}[W|V] + R^{\mathsf{A}}[W|V]\big] \\ &= \mathbb{C}\mathrm{ov}\big[\mathbb{E}^{\mathsf{A}}[U|V], \mathbb{E}^{\mathsf{A}}[W|V]\big] + 0 + 0 + \mathbb{C}\mathrm{ov}\big[R^{\mathsf{A}}[U|V], R^{\mathsf{A}}[W|V]\big] \\ &= \mathbb{C}\mathrm{ov}\big[\mathbb{E}^{\mathsf{A}}[U|V], \mathbb{E}^{\mathsf{A}}[W|V]\big] + \mathbb{E}\big[\mathbb{C}\mathrm{ov}^{\mathsf{A}}[U, W|V]\big],\end{aligned}$$

proving the law of total linear covariance in (g). The inequality $\mathbb{C}\mathrm{ov}[U] \geq \mathbb{C}\mathrm{ov}[\mathbb{E}[U|V]]$ is well known (it follows from the common law of total covariance). For the second inequality, let $U' := \mathbb{E}[U|V]$. By the first tower property in (f), $\mathbb{E}^{\mathsf{A}}[U'|V] = \mathbb{E}^{\mathsf{A}}[U|V]$. Hence, by the law of total linear covariance that was just established,

$$\begin{aligned}\mathbb{C}\mathrm{ov}\big[U'\big] &= \mathbb{C}\mathrm{ov}\big[\mathbb{E}^{\mathsf{A}}\big[U'|V\big]\big] + \mathbb{E}\big[\mathbb{C}\mathrm{ov}^{\mathsf{A}}\big[U'|V\big]\big] = \mathbb{C}\mathrm{ov}\big[\mathbb{E}^{\mathsf{A}}[U|V]\big] + \mathbb{C}\mathrm{ov}\big[R^{\mathsf{A}}\big[U'|V\big]\big] \\ &\geq \mathbb{C}\mathrm{ov}\big[\mathbb{E}^{\mathsf{A}}[U|V]\big],\end{aligned}$$

where we used Lemma 4.4 and the law of total linear expectation (d) in the second step, finalising the proof of (g).

In order to prove (h), note that the independence of W and (U, V) implies

$$\begin{aligned}\mathbb{C}\mathrm{ov}[W \otimes U, V] &= \mathbb{E}[W \otimes U \otimes V] - \mathbb{E}[W \otimes U] \otimes \mu_V = \mu_W \otimes \mathbb{E}[U \otimes V] - \mu_W \otimes \mu_U \otimes \mu_V \\ &= \mu_W \otimes C_{UV}.\end{aligned}$$

Further, by Lemma 4.4, $\mathbb{C}\mathrm{ov}[\mathbb{E}^{\mathsf{A}}[U|V], V] = C_{UV}$. Since $\mathbb{E}[\mu_W \otimes \mathbb{E}^{\mathsf{A}}[U|V]] = \mu_W \otimes \mu_U = \mathbb{E}[W \otimes U]$, Lemma A.6 implies that, for any $\gamma \in \mathsf{A}(\mathcal{H}; \mathcal{F} \otimes \mathcal{G})$,

$$\begin{aligned}\big\langle W \otimes U - \mu_W \otimes \mathbb{E}^{\mathsf{A}}[U|V], \gamma \circ V\big\rangle_{L^2(\mathbb{P}; \mathcal{F}\otimes\mathcal{G})} &= \mathrm{tr}\big(\mathbb{C}\mathrm{ov}\big[W \otimes U - \mu_W \otimes \mathbb{E}^{\mathsf{A}}[U|V], V\big]\overline{\gamma}^*\big) \\ &= \mathrm{tr}\big(\big(\mathbb{C}\mathrm{ov}[W \otimes U, V] - \mu_W \otimes C_{UV}\big)\overline{\gamma}^*\big) \\ &= 0.\end{aligned}$$

Therefore, $\mu_W \otimes \mathbb{E}^{\mathsf{A}}[U|V]$ is the $L^2(\mathbb{P}; \mathcal{F} \otimes \mathcal{G})$-orthogonal projection of $W \otimes U$ onto $\overline{\mathsf{A}_{\mathcal{F}\otimes\mathcal{G}} \circ V}$.

In order to prove[9] (i) first note that, by linearity (b), we may assume that $U = 0$. Further, since $(\alpha) \implies (\beta)$, we may simply assume that (β) holds. This implies that

$$\mu_{U_k} \xrightarrow[k\to\infty]{} 0, \qquad \|C_{U_k}\| \xrightarrow[k\to\infty]{} 0, \qquad C_{U_k V} = C_{U_k}^{1/2} R_{U_k V} C_V^{1/2} \xrightarrow[k\to\infty]{} 0, \tag{8.1}$$

[9] A simpler proof can be obtained from using Theorem 4.8: After establishing (8.1), (i) follows from

$$\mathbb{E}^{\mathsf{A}}[U_k|V] = \mu_{U_k} + \big(C_V^{\dagger} C_{V U_k}\big)^*(V - \mu_V) \xrightarrow[k\to\infty]{\text{a.s.}} 0.$$

where we used Theorem A.3 and adopted the notation therein (note that $C_{U_k V}$ has finite rank, since C_V has finite rank by assumption). By Lemma 4.4 and Lemma A.6,

$$\overline{\gamma}^{\mathsf{A}}_{U_k|V} C_V = \mathbb{C}\mathrm{ov}\big[\gamma^{\mathsf{A}}_{U_k|V} \circ V, V\big] = \mathbb{C}\mathrm{ov}\big[\mathbb{E}^{\mathsf{A}}[U_k|V], V\big] = C_{U_k V} \xrightarrow[k\to\infty]{} 0.$$

Therefore, $\overline{\gamma}^{\mathsf{A}}_{U_k|V}|_{\operatorname{ran} C_V} \xrightarrow[k\to\infty]{} 0$ and, by the assumption of finite rank,

$$V \in \operatorname{ran} C_V \quad \text{a.s.} \quad \text{and} \quad \overline{\gamma}^{\mathsf{A}}_{U_k|V} \circ V \xrightarrow[k\to\infty]{\text{a.s.}} 0.$$

Denoting the constant part of $\gamma^{\mathsf{A}}_{U_k|V}$ by $b_k \in \mathcal{G}$, that is, $\gamma^{\mathsf{A}}_{U_k|V}(v) = b_k + \overline{\gamma}^{\mathsf{A}}_{U_k|V}(v)$ for $v \in \mathcal{H}$, the law of total linear expectation (d) implies

$$b_k + \overline{\gamma}^{\mathsf{A}}_{U_k|V} \mu_V = \mathbb{E}\big[\mathbb{E}^{\mathsf{A}}[U_k|V]\big] = \mu_{U_k} \xrightarrow[k\to\infty]{\text{a.s.}} 0.$$

Since $\mu_V \in \operatorname{ran} C_V$ and $\overline{\gamma}^{\mathsf{A}}_{U_k|V}|_{\operatorname{ran} C_V} \xrightarrow[k\to\infty]{} 0$, we obtain $b_k \xrightarrow[k\to\infty]{} 0$ and thereby

$$\mathbb{E}^{\mathsf{A}}[U_k|V] = b_k + \overline{\gamma}^{\mathsf{A}}_{U_k|V} \circ V \xrightarrow[k\to\infty]{\text{a.s.}} 0. \qquad \square$$

Proof of Theorem 4.7. We choose $\mathcal{H} = \mathcal{G} = \mathcal{F} = \mathbb{R}$ for all counterexamples provided in this proof. For counterexamples to (a)–(e), let $\mathbb{P}$ to be the uniform distribution on $\Omega = \{1, 2, 3\}$ and the random variables V, $U_1 := V$ and $U_2 := W := |V|$ be given by

$\omega \in \Omega$	$U_1(\omega) = V(\omega)$	$U_2(\omega) = W = \lvert V(\omega)\rvert$	$\mathbb{E}^{\mathsf{A}}[U_1\lvert V](\omega)$	$\mathbb{E}^{\mathsf{A}}[U_2\lvert V](\omega)$
1	-1	1	-1	$\frac{2}{3}$
2	0	0	0	$\frac{2}{3}$
3	1	1	1	$\frac{2}{3}$

Clearly, $\mathbb{E}^{\mathsf{A}}[U_1|V] = U_1 = V$ and, by solving a simple linear regression (or simply by symmetry and Theorem 4.5(d)), $\mathbb{E}^{\mathsf{A}}[U_2|V] \equiv \frac{2}{3}$, as illustrated in Figure 4 (left). Therefore, $U_2 \geq U_1$, but $\mathbb{E}^{\mathsf{A}}[U_2|V] \not\geq \mathbb{E}^{\mathsf{A}}[U_1|V]$, disproving (a). Further, $|\mathbb{E}^{\mathsf{A}}[U_1|V]| \leq \mathbb{E}^{\mathsf{A}}[|U_1||V]$ does not hold, providing a counterexample to (b) and (c). For $f\colon \mathbb{R} \to \mathbb{R}$, $f(x) = x$, we obtain $f(V)\mathbb{E}^{\mathsf{A}}[U_1|V] = V^2$, which clearly cannot equal $\mathbb{E}^{\mathsf{A}}[f(V)U_1|V]$, since it is not an affine transformation of V, disproving (d). Finally, $\mathbb{E}^{\mathsf{A}}[\mathbb{E}^{\mathsf{A}}[U_2|V] \mid W] = \mathbb{E}^{\mathsf{A}}[\frac{2}{3}|W] = \frac{2}{3}$ a.s., which clearly differs from $\mathbb{E}^{\mathsf{A}}[U_2|W] = W$ and thereby provides a counterexample to (e).

Since $\mathbb{E}^{\mathsf{A}}$ lacks monotonicity, a counterexample to (f) is easy to construct. Consider the uniform distribution $\mathbb{P}$ on $\Omega = \{1, 2, 3\}$ and V as well as the sequence $(U_k)_{k\in\mathbb{N}}$ given by

$\omega \in \Omega$	$V(\omega)$	$U_{2k+1}(\omega)$	$U_{2k}(\omega)$	$\liminf_{k\to\infty} U_k(\omega)$	$\liminf_{k\to\infty} \mathbb{E}^{\mathsf{A}}[U_k\lvert V](\omega)$
1	-1	0	1	0	$-\frac{1}{6}$
2	0	0	0	0	$\frac{2}{6}$
3	1	1	0	0	$-\frac{1}{6}$

Then $\mathbb{E}^{\mathsf{A}}[\liminf_{k\to\infty} U_k|V] = \mathbb{E}^{\mathsf{A}}[0|V] = 0$, while $\liminf_{k\to\infty} \mathbb{E}^{\mathsf{A}}[U_k|V](\omega) < 0$ for $\omega = 1$ and $\omega = 3$, which follows from the solution of a simple linear regression problem and is visualised in Figure 4(right).

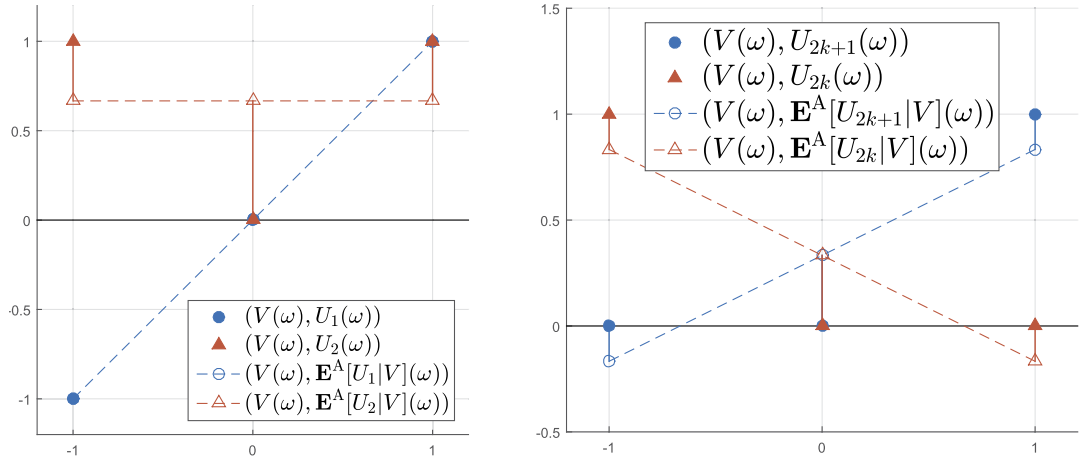

Figure 4. *Left*: Counterexample to Theorem 4.7(a)–(e) as described above, with the corresponding LCEs $\mathbb{E}^{\mathsf{A}}[U_k|V]$, $k = 1, 2$. *Right*: Counterexample to Fatou's lemma (Theorem 4.7(f)) with corresponding LCEs $\mathbb{E}^{\mathsf{A}}[U_k|V]$, $k \in \mathbb{N}$.

Let us now construct a counterexample to the (conventional) dominated convergence theorem (g). Let $\varepsilon > 0$ and $\alpha = (2+2\varepsilon)^{-1}$, for example, $\varepsilon = 1/4$ and $\alpha = 2/5$. Let $\mathbb{P}$ be the uniform distribution on $\Omega = [-1, 1]$ and

$$V(\omega) = \begin{cases} (1+\omega)^{-\alpha} - 1 & \text{for } \omega \in [-1, 0], \\ -(1-\omega)^{-\alpha} + 1 & \text{for } \omega \in [0, 1], \end{cases}$$

$$U_k(\omega) = \begin{cases} (1+\omega)^{-2\alpha} - 1 & \text{for } \omega \in \left[\frac{1}{2k} - 1, \frac{1}{k} - 1\right], \\ -(1-\omega)^{-2\alpha} + 1 & \text{for } \omega \in \left[1 - \frac{1}{k}, 1 - \frac{1}{2k}\right], \\ 0 & \text{otherwise,} \end{cases}$$

as illustrated in Figure 5. Clearly, each U_k is bounded and thereby lies in $L^2(\mathbb{P}; \mathcal{G})$ and $U_k \xrightarrow[k\to\infty]{\text{a.s.}} 0$. Then, $\mathbb{E}^{\mathsf{A}}[U_k|V] = a_k V + b_k$ for some $a_k, b_k \in \mathbb{R}$ where $b_k = 0$ for symmetry reasons. Let $\beta := 3\alpha - 1$ and note that $\beta > 0$ for sufficiently small ε ($\beta = 1/5$ in the above example). A straightforward computation shows that

$$\begin{aligned} \|a_k V - U_k\|^2_{L^2(\mathbb{P};\mathcal{G})} &= a_k^2 \|V\|^2_{L^2(\mathbb{P};\mathcal{G})} - 2a_k \langle V, U_k \rangle_{L^2(\mathbb{P};\mathcal{G})} + \|U_k\|^2_{L^2(\mathbb{P};\mathcal{G})} \\ &= \frac{2(1+\varepsilon)}{\varepsilon} a_k^2 - \frac{4k^\beta}{\beta}\big(2^\beta - 1\big)a_k + \|U_k\|^2_{L^2(\mathbb{P};\mathcal{G})}, \end{aligned}$$

which is minimised by $a_k = \frac{(2^\beta - 1)\varepsilon}{\beta(1+\varepsilon)} k^\beta \xrightarrow[k\to\infty]{} \infty$, which contradicts $\mathbb{E}^{\mathsf{A}}[U_k|V] = a_k V \xrightarrow[k\to\infty]{\text{a.s.}} 0$.

The following two counterexamples disprove (h) for every $p < 2$ and for every $p > 2$, respectively, (for $p = 2$ the projection is clearly contractive, since it is orthogonal). Choose $\mathbb{P}$ to be the uniform distribution on $\Omega = \{1, 2, 3, 4\}$ and the random variables V, U_1 and U_2 in the following way, where

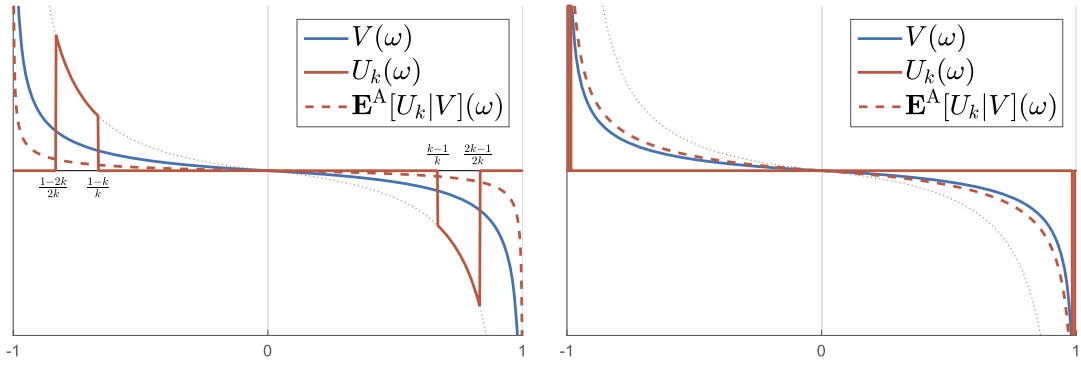

Figure 5. Counterexample to the dominated convergence theorem (Theorem 4.7(g)). Note that we plot $\mathbb{E}^{\mathsf{A}}[U_k|V]$ as a function of ω (and not of V) which is why it not a linear function in contrast to the other plots, while being a multiple of V by the factor α_k. For sufficiently small $\varepsilon > 0$, this factor α_k increases with k (here $k = 3$ (left), $k = 70$ (right) and $\varepsilon = 0.01$) as can be seen from the dotted red lines in the two plots. Therefore, in contrast to $(U_k)_{k\in\mathbb{N}}$, the sequence of LCEs $(\mathbb{E}^{\mathsf{A}}[U_k|V])_{k\in\mathbb{N}}$ does not converge to zero a.s.

$\varepsilon \in (0, 1)$ is a free parameter yet to be chosen.

$\omega \in \Omega$	$V(\omega)$	$U_1(\omega)$	$U_2(\omega)$
1	-1	-1	-1
2	$-\varepsilon$	0	-2ε
3	ε	0	2ε
4	1	1	1

Again, the computation of $\mathbb{E}^{\mathsf{A}}[U_j|V] = a_j V + b_j, a_j, b_j \in \mathbb{R}$, $j = 1, 2$, reduces to a linear regression that is solved by

$$b_1 = b_2 = 0, \qquad a_1 = \frac{1}{1+\varepsilon^2}, \qquad a_2 = \frac{1+2\varepsilon^2}{1+\varepsilon^2}.$$

Note that $\mathbb{E}[|U_1|^p] = \frac{1}{2}$, $\mathbb{E}[|U_2|^p] = \frac{1}{2}(1 + (2\varepsilon)^p)$ and $\mathbb{E}[|\mathbb{E}^{\mathsf{A}}[U_j|V]|^p] = \frac{1}{2}a_j^p(1 + \varepsilon^p)$, $j = 1, 2$. It follows that the inequality in (h) for $U = U_j$, $j = 1, 2$, holds whenever

$$f_1(\varepsilon) := \left(1+\varepsilon^2\right)^p - \left(1+\varepsilon^p\right) \geq 0, \tag{8.2}$$

$$f_2(\varepsilon) := \left(1+\varepsilon^2\right)^p \left(1+2^p\varepsilon^p\right) - \left(1+2\varepsilon^2\right)^p \left(1+\varepsilon^p\right) \geq 0, \tag{8.3}$$

respectively. Bernoulli's inequality $(1+x)^r \leq 1+rx$ for $x \geq -1$ and exponents $0 \leq r \leq 1$ implies that, for $1 \leq p \leq 2$,

$$\begin{aligned} f_1(\varepsilon) &= \left(1+\varepsilon^2\right)\left(1+\varepsilon^2\right)^{p-1} - \left(1+\varepsilon^p\right) \\ &\leq \left(1+\varepsilon^2\right)\left(1+(p-1)\varepsilon^2\right) - \left(1+\varepsilon^p\right) \\ &= p\varepsilon^2 + (p-1)\varepsilon^4 - \varepsilon^p. \end{aligned}$$

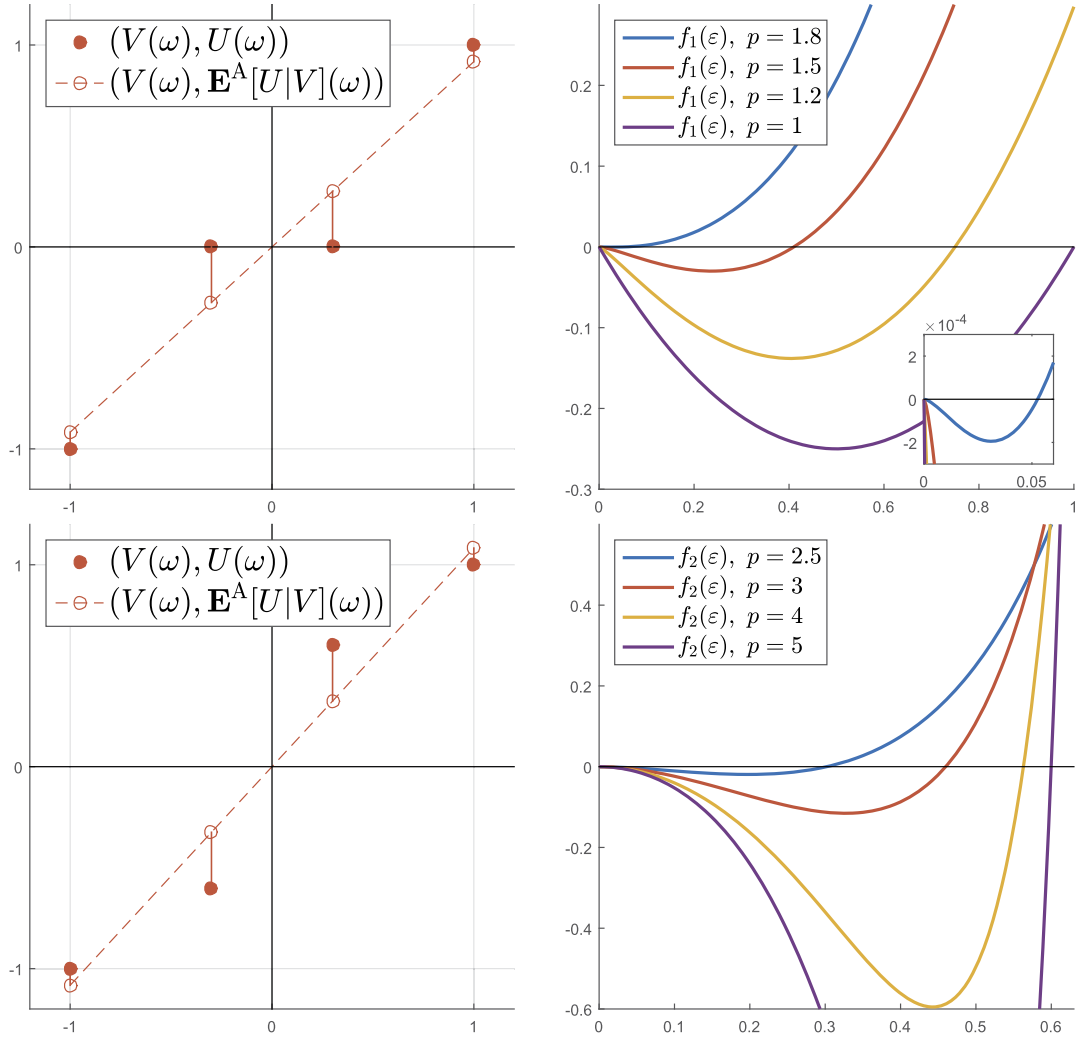

Figure 6. *Top*: Counterexample to Theorem 4.7(h) for $1 \le p < 2$. *Bottom*: Counterexample to Theorem 4.7(h) for $p > 2$. *Left*: LCEs for the above examples and $\varepsilon = 0.3$. *Right*: The functions $f_1(\varepsilon)$ and $f_2(\varepsilon)$ for several values of p. For every $p < 2$ (respectively, $p > 2$) there exists a sufficiently small $\varepsilon > 0$ such that $f_j(\varepsilon) < 0$, $j = 1, 2$.

Since, for any $p < 2$, $p\varepsilon^2 + (p-1)\varepsilon^4 < \varepsilon^p$ for sufficiently small ε, we can falsify (8.2) and thereby disprove (h) for any $p < 2$.

For fixed $p > 2$ consider the Taylor polynomial of degree 2 for f_2, namely $T_2 f_2(\varepsilon) = -p\varepsilon^2$; note that this is *not* a Taylor polynomial of f_2 for $p < 2$. Hence, (8.3) cannot hold for sufficiently small ε, providing a counterexample to (h) for any $p > 2$.

Figure 6 illustrates the counterexamples for $\varepsilon = 0.3$ (left) as well as the functions $f_1(\varepsilon)$ and $f_2(\varepsilon)$ for several values of p.

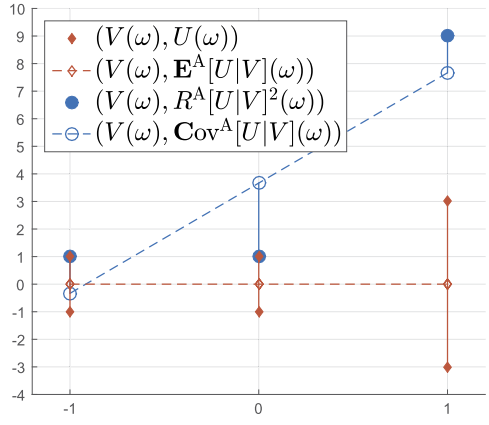

Figure 7. In contrast to the conditional covariance $\mathbb{C}\mathrm{ov}[U|V]$, the LCC $\mathbb{C}\mathrm{ov}^{\mathsf{A}}[U|V]$ can take on negative values, while its expected value $\mathbb{C}\mathrm{ov}_V^{\mathsf{A}}[U]$ is guaranteed to be non-negative.

We now give a counterexample to (i). Let $\mathbb{P}$ be the uniform distribution on $\Omega = \{1, \dots, 6\}$ and V and U given by

$\omega \in \Omega$	$V(\omega)$	$U(\omega)$	$\mathbb{E}^{\mathsf{A}}[U\|V](\omega)$	$R^{\mathsf{A}}[U\|V]^2(\omega)$	$\mathbb{C}\mathrm{ov}^{\mathsf{A}}[U\|V](\omega)$
1	-1	1	0	1	$\frac{1}{6}(7-N^2)$
2	-1	-1			
3	0	1	0	1	$\frac{1}{3}(2+N^2)$
4	0	-1			
5	1	N	0	N^2	$\frac{1}{6}(1+5N^2)$
6	1	$-N$			

where $N > 0$. By symmetry, $\mathbb{E}^{\mathsf{A}}[U|V] = \mathbb{E}[U|V] = 0$ while $\mathbb{C}\mathrm{ov}^{\mathsf{A}}[U|V] = \frac{1}{2}(N^2-1)V + \frac{1}{3}(N^2+2)$, which follows from the solution of a simple linear regression problem and is visualised in Figure 7. If $N > 0$ is sufficiently large, $\mathbb{C}\mathrm{ov}^{\mathsf{A}}[U|V](\omega)$ clearly takes on negative values for $\omega = 1, 2$. □

Proof of Theorem 4.8. First note that, by Theorem A.1, $C_V^\dagger C_{VU} \in \mathsf{L}(\mathcal{G}; \mathcal{H})$ is well-defined and bounded and that $C_V(C_V^\dagger C_{VU}) = C_{VU}$, which implies $\overline{\gamma}_{U|V}^{\mathsf{A}} C_V = (C_V^\dagger C_{VU})^* C_V = C_{UV}$. We have to show that $U - \gamma_{U|V}^{\mathsf{A}} \circ V$ is $L^2(\mathbb{P}; \mathcal{G})$-perpendicular to $\gamma \circ V$ for any other $\gamma \in \mathsf{A}(\mathcal{H}; \mathcal{G})$. Since $\mathbb{E}[\gamma_{U|V}^{\mathsf{A}} \circ V] = \mu_U = \mathbb{E}[U]$, it follows by Lemma A.6 that

$$\langle U - \gamma_{U|V}^{\mathsf{A}} \circ V, \gamma \circ V \rangle_{L^2(\mathbb{P};\mathcal{G})} = \mathrm{tr}\big(\mathbb{C}\mathrm{ov}\big[U - \gamma_{U|V}^{\mathsf{A}} \circ V, V\big]\overline{\gamma}^*\big) = \mathrm{tr}\big(\big(C_{UV} - \overline{\gamma}_{U|V}^{\mathsf{A}} C_V\big)\overline{\gamma}^*\big) = 0,$$

as required. □

Proof of Lemma 4.11. The Karhunen–Loève expansion of V takes the form

$$V = \mu_V + \sum_{i \in \mathbb{N}} Z_i h_i,$$

where the Z_i are uncorrelated real-valued random variables over $(\Omega, \Sigma, \mathbb{P})$ with $\mathbb{E}[Z_i] = 0$ and $\mathbb{V}[Z_i] = \sigma_i^2$ for all $i \in \mathbb{N}$. Observe that $\sigma(V^{(n)}) = \sigma(Z^{(n)})$, where $Z^{(n)} := (Z_1, \dots, Z_n)$.

Now let $W \in \overline{\mathsf{A}_{\mathcal{G}} \circ V} \cap L^2(\Omega, \sigma(V^{(n)}); \mathcal{G})$. Since $W \in \overline{\mathsf{A}_{\mathcal{G}} \circ V}$, there exists a sequence $(\gamma_k)_{k\in\mathbb{N}}$ in $\mathsf{A}(\mathcal{H}; \mathcal{G})$ such that $\|\gamma_k \circ V - W\|_{L^2(\mathbb{P};\mathcal{G})} \xrightarrow[k\to\infty]{} 0$. In order to show that $W \in \overline{\mathsf{A}_{\mathcal{G}} \circ V^{(n)}}$, we will find a sequence $(\tilde{\gamma}_k)_{k\in\mathbb{N}}$ in $\mathsf{A}(\mathcal{H}; \mathcal{G})$ such that $\|\tilde{\gamma}_k \circ V^{(n)} - W\|_{L^2(\mathbb{P};\mathcal{G})} \xrightarrow[k\to\infty]{} 0$. To this end, we shift each γ_k by a constant, choosing $\tilde{\gamma}_k(v) := \gamma_k(v) + \gamma_k(\mu_V - \mu_V^{(n)})$, where $\mu_V^{(n)} := \mathbb{E}[V^{(n)}] = P_{\mathcal{H}^{(n)}}\mu_V$. In order to prove above convergence observe that $\mathbb{E}[W|V^{(n)}] = W$, since W is $\sigma(V^{(n)})$-measurable, and that

$$\mathbb{E}\big[\gamma_k(V - \mu_V)|V^{(n)}\big] = \gamma_k\left(\mathbb{E}\left[\sum_{i\in\mathbb{N}} Z_i h_i | Z^{(n)}\right]\right) = \gamma_k\left(\sum_{i=1}^{n} Z_i h_i\right) = \gamma_k\big(V^{(n)} - \mu_V^{(n)}\big),$$

as the random variables Z_i are uncorrelated. Since conditional expectations are $L^2(\mathbb{P}; \mathcal{G})$-contractive projections, it follows that

$$\begin{aligned}
\big\|\tilde{\gamma}_k \circ V^{(n)} - W\big\|_{L^2(\mathbb{P};\mathcal{G})} &= \big\|\gamma_k\big(V^{(n)} - \mu_V^{(n)}\big) - W + \gamma_k(\mu_V)\big\|_{L^2(\mathbb{P};\mathcal{G})} \\
&= \big\|\mathbb{E}\big[\gamma_k(V - \mu_V) \mid V^{(n)}\big] - \mathbb{E}\big[W \mid V^{(n)}\big] + \gamma_k(\mu_V)\big\|_{L^2(\mathbb{P};\mathcal{G})} \\
&= \big\|\mathbb{E}\big[\gamma_k(V) - W|V^{(n)}\big]\big\|_{L^2(\mathbb{P};\mathcal{G})} \\
&\le \|\gamma_k \circ V - W\|_{L^2(\mathbb{P};\mathcal{G})} \\
&\xrightarrow[k\to\infty]{} 0.
\end{aligned}$$

The second inclusion in (4.3) is trivial. □

Proof of Theorem 4.12. Claim (a) follows from Lemma 4.11 via

$$\mathbb{E}\big[\mathbb{E}^{\mathsf{A}}[U|V] \mid V^{(n)}\big] = P_{L^2(\Omega,\sigma(V^{(n)});\mathcal{G})} P_{\overline{\mathsf{A}_{\mathcal{G}}\circ V}} U = P_{\overline{\mathsf{A}_{\mathcal{G}}\circ V^{(n)}}} U = \mathbb{E}^{\mathsf{A}}\big[U|V^{(n)}\big],$$

where all orthogonal projections are taken with respect to the $L^2(\Omega, \Sigma, \mathbb{P}; \mathcal{G})$ inner product.

Since $\mathbb{E}^{\mathsf{A}}[U|V] = \mathbb{E}[\mathbb{E}^{\mathsf{A}}[U|V] \mid V]$ by Theorem 4.5(e) and using (4.4), claim (b) follows directly from [5], Theorems 2 and 6, or [23], Theorems 11.7 and 11.10. □

Proof of Theorem 4.13. Since $C_V^{(n)}$ has finite rank, Theorem A.3 yields $\operatorname{ran} C_{VU}^{(n)} \subseteq \operatorname{ran} C_V^{(n)}$ and Theorem 4.8 implies

$$\mathbb{E}^{\mathsf{A}}\big[U|V^{(n)}\big] = P_{\overline{\mathsf{A}_{\mathcal{G}}\circ V^{(n)}}} U = \gamma_{U|V}^{(n)} \circ V^{(n)} = \gamma_{U|V}^{(n)} \circ V.$$

The statements follow directly from Theorem 4.12. □

Proof of Theorem 4.14. First note that $\gamma_\varepsilon^{\mathsf{A}_2} \in \mathsf{A}_2(\mathcal{H}; \mathcal{G})$, since it is a shifted composition of the Hilbert–Schmidt operator C_{UV} and the bounded operator $(C_V + \varepsilon \operatorname{Id}_{\mathcal{H}})^{-1}$. Now let $\gamma \in \mathsf{A}_2(\mathcal{H}; \mathcal{G})$ and $\delta := \gamma - \gamma_\varepsilon^{\mathsf{A}_2} \in \mathsf{A}_2(\mathcal{H}; \mathcal{G})$. Since $\mathbb{E}[\gamma_\varepsilon^{\mathsf{A}_2} \circ V] = \mu_U$, Lemma A.6 implies that

$$\begin{aligned}
\mathbb{E}\big[\big\langle U - \gamma_\varepsilon^{\mathsf{A}_2} \circ V, \delta \circ V\big\rangle_{\mathcal{G}}\big] &= \operatorname{tr}\big(C_{UV}\overline{\delta}^* - \overline{\gamma}_\varepsilon^{\mathsf{A}_2} C_V \overline{\delta}^*\big) \\
&= \operatorname{tr}\big(C_{UV}(C_V + \varepsilon \operatorname{Id}_{\mathcal{H}})^{-1}(C_V + \varepsilon \operatorname{Id}_{\mathcal{H}} - C_V)\overline{\delta}^*\big) \\
&= \varepsilon \operatorname{tr}\big(\overline{\gamma}_\varepsilon^{\mathsf{A}_2}\overline{\delta}^*\big) \\
&= \varepsilon\big\langle \gamma_\varepsilon^{\mathsf{A}_2}, \delta\big\rangle_{\mathsf{L}_2}.
\end{aligned}$$

Hence,

$$\mathcal{E}^{\mathrm{reg}}_{U|V}(\gamma) = \mathcal{E}^{\mathrm{reg}}_{U|V}\big(\gamma_\varepsilon^{\mathsf{A}_2}\big) + \mathbb{E}\big[\big\|\delta(V)\big\|_{\mathcal{G}}^2\big] + \varepsilon\|\delta\|_{\mathsf{L}_2}^2 \underbrace{-2\mathbb{E}\big[\big\langle U - \gamma_\varepsilon^{\mathsf{A}_2}(V), \delta(V)\big\rangle_{\mathcal{G}}\big] + 2\varepsilon\big\langle\gamma_\varepsilon^{\mathsf{A}_2}, \delta\big\rangle_{\mathsf{L}_2}}_{=0}$$

$$\geq \mathcal{E}^{\mathrm{reg}}_{U|V}\big(\gamma_\varepsilon^{\mathsf{A}_2}\big),$$

proving the claim. □

Proof of Theorem 4.15. By the law of total linear expectation in Theorem 4.5(d),

$$\begin{aligned}\mathbb{E}\big[\mathbb{C}\mathrm{ov}^{\mathsf{A}}[U, W|V]\big] &= \mathbb{E}\big[\mathbb{E}^{\mathsf{A}}\big[R^{\mathsf{A}}[U|V] \otimes R^{\mathsf{A}}[W|V] \mid V\big]\big] \\ &= \mathbb{E}\big[R^{\mathsf{A}}[U|V] \otimes R^{\mathsf{A}}[W|V]\big] \\ &= \mathbb{C}\mathrm{ov}_V^{\mathsf{A}}[U, W],\end{aligned}$$

proving (a). By the law of total covariance and its linear version in Theorem 4.5(g), we obtain

$$\mathbb{E}\big[\mathbb{C}\mathrm{ov}[U|V]\big] = \mathbb{C}\mathrm{ov}[U] - \mathbb{C}\mathrm{ov}\big[\mathbb{E}[U|V]\big] \leq \mathbb{C}\mathrm{ov}[U] - \mathbb{C}\mathrm{ov}\big[\mathbb{E}^{\mathsf{A}}[U|V]\big] = \mathbb{E}\big[\mathbb{C}\mathrm{ov}^{\mathsf{A}}[U|V]\big],$$

proving (b) (the equality in (b) follows directly from (a)). In order to prove (c), first note that, by Theorem A.3 and using Notation 4.10,

$$C_V^{1/2}\big(\overline{\gamma}^{(n)}_{U|V}\big)^* = C_V^{1/2} C_V^{(n)\dagger} C_{VU}^{(n)} = P_{\mathcal{H}^{(n)}} R_{VU} C_U^{1/2} \xrightarrow[n\to\infty]{} R_{VU} C_U^{1/2} = M_{VU}.$$

Hence, by (4.6) and Lemma A.6,

$$\begin{aligned}\mathbb{C}\mathrm{ov}\big[\mathbb{E}^{\mathsf{A}}[U|V], \mathbb{E}^{\mathsf{A}}[W|V]\big] &= \lim_{n\to\infty} \mathbb{C}\mathrm{ov}\big[\gamma^{(n)}_{U|V} \circ V, \gamma^{(n)}_{W|V} \circ V\big] \\ &= \lim_{n\to\infty} \overline{\gamma}^{(n)}_{U|V} C_V \big(\overline{\gamma}^{(n)}_{W|V}\big)^* \\ &= \lim_{n\to\infty} \big(C_V^{1/2}\big(\overline{\gamma}^{(n)}_{U|V}\big)^*\big)^* \big(C_V^{1/2}\big(\overline{\gamma}^{(n)}_{W|V}\big)^*\big) \\ &= M_{VU}^* M_{VW}.\end{aligned}$$

By (a) and the law of total linear covariance in Theorem 4.5(g), we obtain

$$\mathbb{C}\mathrm{ov}_V^{\mathsf{A}}[U, W] = \mathbb{C}\mathrm{ov}[U, W] - \mathbb{C}\mathrm{ov}\big[\mathbb{E}^{\mathsf{A}}[U|V], \mathbb{E}^{\mathsf{A}}[W|V]\big] = C_{UW} - M_{VU}^* M_{VW},$$

thus completing the proof. □

Proof of Corollary 4.18. Noting that $\mu_Z = \mathbb{C}\mathrm{ov}_V^{\mathsf{A}}[U, W]$, the claim follows directly from Theorems 4.8, 4.13, and 4.15. □

Proof of Proposition 5.6. In this proof $\mathfrak{m}$ will be viewed as an element of $\mathsf{L}_2(\mathcal{G}; L^2(\mathbb{P}_X))$, which is isometrically isomorphic to $\mathcal{G} \otimes L^2(\mathbb{P}_X) \cong L^2(\mathbb{P}_X; \mathcal{G})$ (see Remark 5.3). In this case, $f_g = \mathfrak{m}(g)$ by (5.3).

If $\mathfrak{m} \in \mathsf{L}_2(\mathcal{G}; \mathcal{H})$, then clearly $f_g = \mathfrak{m}(g) \in \mathcal{H}$ and this shows that (A) $\implies$ ($\mathrm{A}_{\mathrm{old}}$).

Now let $[\mathfrak{m}] \in (\mathcal{G} \otimes \mathcal{H})_{\mathcal{C}}$. Then there exist $\mathfrak{h} \in \mathcal{G} \otimes \mathcal{H}$ and $c \in \mathcal{G}$ such that $\mathfrak{h}(x) + c = \mathfrak{m}(x)$ for $\mathbb{P}_X$-a.e. $x \in \mathcal{X}$, which implies $f_g = \mathfrak{m}(g) = \mathfrak{h}(g) + \langle c, g\rangle_{\mathcal{G}}$ $\mathbb{P}_X$-a.e. in $\mathcal{X}$. Since $\mathfrak{h}(g) \in \mathcal{H}$ and $\langle c, g\rangle_{\mathcal{G}} \in \mathbb{R}$ for each $g \in \mathcal{G}$, this shows that (B) $\implies$ ($\mathrm{B}_{\mathrm{old}}$).

Let $\mathfrak{h} \in \mathcal{G} \otimes \mathcal{H}$ be such that $[\mathfrak{h}] = P_{\overline{(\mathcal{G}\otimes\mathcal{H})_{\mathcal{C}}}^{L^2_{\mathcal{C}}(\mathbb{P}_X;\mathcal{G})}}[\mathfrak{m}]$. Letting $c := \mathbb{E}[(\mathfrak{m} - \mathfrak{h})(X)] \in \mathcal{G}$ and denoting the unit constant function by $\mathbb{1} \in L^2(\mathbb{P}_X)$, it follows that, for each $h \in \mathcal{H}$ and $g \in \mathcal{G}$,

$$\begin{aligned} 0 &= \big\langle [\mathfrak{h}] - [\mathfrak{m}], [g \otimes h] \big\rangle_{L^2_{\mathcal{C}}(\mathbb{P}_X;\mathcal{G})} \\ &= \big\langle \mathfrak{h} + c \otimes \mathbb{1} - \mathfrak{m}, g \otimes h - g \otimes \mathbb{E}\big[h(X)\big] \big\rangle_{\mathcal{G}\otimes L^2(\mathbb{P}_X)} \\ &= \big\langle \mathfrak{h}(g) + \langle c, g\rangle_{\mathcal{G}}\mathbb{1} - \mathfrak{m}(g), h - \mathbb{E}\big[h(X)\big] \big\rangle_{L^2(\mathbb{P}_X)} \\ &= \big\langle \big[\mathfrak{h}(g)\big] - \big[\mathfrak{m}(g)\big], [h] \big\rangle_{L^2_{\mathcal{C}}(\mathbb{P}_X)}, \end{aligned}$$

where we used (5.3) and $\langle c, g\rangle_{\mathcal{G}} = \mathbb{E}[(\mathfrak{m}(g) - \mathfrak{h}(g))(X)]$. Since $\mathfrak{h}(g) \in \mathcal{H}$, this shows that (C) $\implies$ ($\mathrm{C}_{\mathrm{old}}$).

If $\mathfrak{h} := P_{\overline{\mathcal{G}\otimes\mathcal{H}}^{L^2(\mathbb{P}_X;\mathcal{G})}}\mathfrak{m} \in \mathcal{G} \otimes \mathcal{H}$, then, for each $h \in \mathcal{H}$ and $g \in \mathcal{G}$,

$$0 = \langle \mathfrak{h} - \mathfrak{m}, g \otimes h\rangle_{\mathcal{G}\otimes L^2(\mathbb{P}_X)} = \big\langle \mathfrak{h}(g) - \mathfrak{m}(g), h \big\rangle_{L^2(\mathbb{P}_X)},$$

where we used (5.3). Since $\mathfrak{h}(g) \in \mathcal{H}$, this shows that (^{u}C) $\implies$ (${}^u\mathrm{C}_{\mathrm{old}}$).

If $\mathfrak{m} \in \overline{\mathcal{G} \otimes \mathcal{H}}^{L^2(\mathbb{P}_X;\mathcal{G})}$, then there exists a sequence $(\mathfrak{h}_n)_{n\in\mathbb{N}}$ in $\mathcal{G} \otimes \mathcal{H}$ such that $\|\mathfrak{h}_n - \mathfrak{m}\|_{\mathsf{L}_2(\mathcal{G};L^2(\mathbb{P}_X))} \to 0$ as $n \to \infty$. Let $g \in \mathcal{G}$ and $h_n := \mathfrak{h}_n(g) \in \mathcal{H}$, $n \in \mathbb{N}$. Then

$$\|h_n - f_g\|_{L^2(\mathbb{P}_X)} = \big\|\mathfrak{h}_n(g) - \mathfrak{m}(g)\big\|_{L^2(\mathbb{P}_X)} \le \|\mathfrak{h}_n - \mathfrak{m}\|_{\mathsf{L}_2(\mathcal{G};L^2(\mathbb{P}_X))}\|g\|_{\mathcal{G}} \xrightarrow[n\to\infty]{} 0,$$

which proves (A*) $\implies$ ($\mathrm{A}^*_{\mathrm{old}}$).

Finally, let $[\mathfrak{m}] \in \overline{(\mathcal{G} \otimes \mathcal{H})_{\mathcal{C}}}^{L^2_{\mathcal{C}}(\mathbb{P}_X;\mathcal{G})}$. Then there exists a sequence $(\mathfrak{h}_n)_{n\in\mathbb{N}}$ in $\mathcal{G} \otimes \mathcal{H}$ such that $\|[\mathfrak{h}_n] - [\mathfrak{m}]\|_{\mathcal{G}\otimes L^2_{\mathcal{C}}(\mathbb{P}_X)} \to 0$ as $n \to \infty$, that is, $\|\mathfrak{h}_n + c_n \otimes \mathbb{1} - \mathfrak{m}\|_{\mathcal{G}\otimes L^2(\mathbb{P}_X)} \to 0$ as $n \to \infty$ where $c_n := \mathbb{E}[(\mathfrak{m} - \mathfrak{h}_n)(X)] \in \mathcal{G}$ and $\mathbb{1} \in L^2(\mathbb{P}_X)$ is the unit constant function. Let $g \in \mathcal{G}$, $h_n := \mathfrak{h}_n(g) \in \mathcal{H}$ and $r_n := \langle c_n, g\rangle_{\mathcal{G}}\mathbb{1}$, $n \in \mathbb{N}$. Then, as $n \to \infty$,

$$\|h_n + r_n - f_g\|_{L^2(\mathbb{P}_X)} = \big\|\mathfrak{h}_n(g) + \langle c_n, g\rangle_{\mathcal{G}}\mathbb{1} - \mathfrak{m}(g)\big\|_{L^2(\mathbb{P}_X)} \le \|\mathfrak{h}_n + c_n \otimes \mathbb{1} - \mathfrak{m}\|_{\mathsf{L}_2(\mathcal{G};L^2(\mathbb{P}_X))}\|g\|_{\mathcal{G}} \to 0,$$

which proves (B*) $\implies$ ($\mathrm{B}^*_{\mathrm{old}}$).

The last two implications follow from the above and [22], Theorems 4.1 and 5.1. □

Proof of Lemma 5.7. Suppose that $(\mathcal{G} \otimes \mathcal{H})_{\mathcal{C}}$ is not dense in $L^2_{\mathcal{C}}(\mathbb{P}_X;\mathcal{G})$. Then there exists $\mathfrak{f} \in L^2(\mathbb{P}_X;\mathcal{G})$ that is not $\mathbb{P}_X$-a.e. constant (i.e., there is no $c \in \mathcal{G}$ such that $\mathfrak{f}(x) = c$ for $\mathbb{P}_X$-a.e. $x \in \mathcal{X}$) such that $[\mathfrak{f}] \perp_{L^2_{\mathcal{C}}(\mathbb{P}_X;\mathcal{G})} (\mathcal{G} \otimes \mathcal{H})_{\mathcal{C}}$. Let $\tilde{\mathfrak{f}} := \mathfrak{f} - \mathbb{E}[\mathfrak{f}(X)]$ and $\mathbb{P}_{\tilde{\mathfrak{f}}} = \tilde{\mathfrak{f}}_{\#}\mathbb{P}_X$ denote the pushforward measure of $\mathbb{P}_X$ under $\tilde{\mathfrak{f}}$.

Then there exists $g_* \in \operatorname{supp}(\mathbb{P}_{\tilde{\mathfrak{f}}}) \subseteq \mathcal{G}$ such that $g_* \neq 0$ (otherwise $\mathbb{P}_{\tilde{\mathfrak{f}}}(\mathcal{G} \setminus \{0\}) = 0$ and therefore $\tilde{\mathfrak{f}} = 0$ $\mathbb{P}_X$-a.e.). Hence, after proper normalisation of $\tilde{\mathfrak{f}}$, we can define the following two *distinct* probability measures on $\mathcal{X}$:

$$Q_1(E) := \int_E \big|\langle \tilde{\mathfrak{f}}(x), g_*\rangle_{\mathcal{G}}\big| \,\mathrm{d}\mathbb{P}_X(x), \qquad Q_2(E) := \int_E \big|\langle \tilde{\mathfrak{f}}(x), g_*\rangle_{\mathcal{G}}\big| - \langle \tilde{\mathfrak{f}}(x), g_*\rangle_{\mathcal{G}} \,\mathrm{d}\mathbb{P}_X(x)$$

for every measurable subset $E \subseteq \mathcal{X}$. Indeed, for $\varepsilon := \|g_*\|_{\mathcal{G}}/2 > 0$ and any $g = g_* + w \in B_\varepsilon(g_*) := \{g_* + w \mid w \in \mathcal{G}, \|w\|_{\mathcal{G}} < \varepsilon\}$, the reverse triangle inequality implies $\langle g, g_*\rangle_{\mathcal{G}} \geq \|g_*\|_{\mathcal{G}}^2 - \|w\|_{\mathcal{G}}\|g_*\|_{\mathcal{G}} > 2\varepsilon^2$. Hence, since $g_* \in \operatorname{supp}(\mathbb{P}_{\tilde{\mathfrak{f}}})$, it follows that, for $E = \tilde{\mathfrak{f}}^{-1}(B_\varepsilon(g_*))$,

$$Q_1(E) - Q_2(E) = \int_E \big\langle \tilde{\mathfrak{f}}(x), g_* \big\rangle_{\mathcal{G}} \,\mathrm{d}\mathbb{P}_X(x) \geq 2\varepsilon^2 \mathbb{P}_X(E) = 2\varepsilon^2 \mathbb{P}_{\tilde{\mathfrak{f}}}\big(B_\varepsilon(g_*)\big) > 0.$$

Since, for every $\mathfrak{h} \in \mathcal{G} \otimes \mathcal{H}$,

$$\langle \tilde{\mathfrak{f}}, \mathfrak{h}\rangle_{L^2(\mathbb{P}_X;\mathcal{G})} = \big\langle \mathfrak{f} - \mathbb{E}\big[\mathfrak{f}(X)\big], \mathfrak{h}\big\rangle_{L^2(\mathbb{P}_X;\mathcal{G})} \overset{[\mathfrak{f}]\perp(\mathcal{G}\otimes\mathcal{H})_C}{=} \big\langle \mathfrak{f} - \mathbb{E}\big[\mathfrak{f}(X)\big], \mathbb{E}\big[\mathfrak{h}(X)\big]\big\rangle_{L^2(\mathbb{P}_X;\mathcal{G})} = 0,$$

it follows that $\tilde{\mathfrak{f}} \perp_{L^2(\mathbb{P}_X;\mathcal{G})} \mathcal{G} \otimes \mathcal{H}$. Let $Z_1 \sim Q_1$, $Z_2 \sim Q_2$ and $x \in \mathcal{X}$. Since $g_* \otimes \varphi(x) \in \mathcal{G} \otimes \mathcal{H}$,

$$\big(\mathbb{E}\big[\varphi(Z_1)\big] - \mathbb{E}\big[\varphi(Z_2)\big]\big)(x) = \int_{\mathcal{X}} k\big(x, x'\big)\big\langle \tilde{\mathfrak{f}}\big(x'\big), g_*\big\rangle_{\mathcal{G}} \,\mathrm{d}\mathbb{P}_X\big(x'\big) = \big\langle g_* \otimes \varphi(x), \tilde{\mathfrak{f}}\big\rangle_{L^2(\mathbb{P}_X;\mathcal{G})} = 0,$$

where we used (5.2), contradicting the assumption of k being characteristic. □

Proof of Theorem 5.8. By Assumption 5.4(B), $\mathfrak{m}(x) = \mathfrak{f}(x) + c$ with $\mathfrak{f} \in \mathcal{G} \otimes \mathcal{H}$ and $c \in \mathcal{G}$. As discussed in Remark 5.3, we can view $\mathfrak{f} \in \mathcal{G} \otimes \mathcal{H}$ as an element of both $L^2(\mathbb{P}_X;\mathcal{G})$ and $\mathsf{L}_2(\mathcal{H};\mathcal{G})$, and thereby $\mathfrak{m}$ as an element of $\mathsf{A}_2(\mathcal{H};\mathcal{G})$. Hence, (5.2) and the injectivity of φ imply that

$$\mathbb{E}[U|V] = \mathbb{E}\big[\psi(Y)|X\big] = \mathfrak{m}(X) = \mathfrak{f}(X) + c = \mathfrak{f}\big(\varphi(X)\big) + c = \mathfrak{m} \circ V \quad \text{a.s.}$$

Since $\mathfrak{m} \in \mathsf{A}_2(\mathcal{H};\mathcal{G}) \subseteq \mathsf{A}(\mathcal{H};\mathcal{G})$, the statements follow from Theorem 4.8; the inclusion $\operatorname{ran} C_{VU} \subseteq \operatorname{ran} C_V$ follows from Assumption 5.4(B) by Proposition 5.6, cf. Figure 3. □

Proof of Theorem 5.11. By Assumption 5.4(B*), there exists a sequence $\mathfrak{f}^{(n)} \in \mathcal{G} \otimes \mathcal{H}$, $n \in \mathbb{N}$, such that

$$\big\|\big[\mathfrak{f}^{(n)}\big] - [\mathfrak{m}]\big\|_{L^2_C(\mathbb{P}_X;\mathcal{G})} \xrightarrow[n\to\infty]{} 0.$$

Therefore, denoting $c^{(n)} := \mathbb{E}[\mathfrak{m}(X) - \mathfrak{f}^{(n)}(X)] \in \mathcal{G}$,

$$\big\|\mathfrak{f}^{(n)}(X) + c^{(n)} - \mathfrak{m}(X)\big\|_{L^2(\mathbb{P};\mathcal{G})} \xrightarrow[n\to\infty]{} 0.$$

As discussed in Remark 5.3, $\mathfrak{f}^{(n)} \in \mathcal{G} \otimes \mathcal{H}$ can be seen as an element of both $L^2(\mathbb{P}_X;\mathcal{G})$ and $\mathsf{L}_2(\mathcal{H};\mathcal{G})$. Hence, (5.2) and the injectivity of φ imply

$$\begin{aligned}\mathbb{E}[U|V] = \mathbb{E}\big[\psi(Y)|X\big] = \mathfrak{m}(X) &= \lim_{n\to\infty} \mathfrak{f}^{(n)}(X) + c^{(n)} \\ &= \lim_{n\to\infty} \mathfrak{f}^{(n)}\big(\varphi(X)\big) + c^{(n)} = \lim_{n\to\infty} \gamma^{(n)} \circ V \quad \text{a.s.},\end{aligned}$$

where the limits are in $L^2(\mathbb{P};\mathcal{G})$ and $\gamma^{(n)}(h) := \mathfrak{f}^{(n)}(h) + c^{(n)}$. Since $\gamma^{(n)} \in \mathsf{A}_2(\mathcal{H};\mathcal{G}) \subseteq \mathsf{A}(\mathcal{H};\mathcal{G})$, this implies that $\mathbb{E}[U|V] \in \overline{\mathsf{A}_{\mathcal{G}} \circ V}$ and thereby $\mathbb{E}[U|V] = \mathbb{E}^{\mathsf{A}}[U|V]$. The claim now follows from Theorem 4.13. □

Appendix: Technical results

The following well-known result due to [9], Theorem 1, (see also [14], Theorem 2.1) is used several times.

Theorem A.1. *Let $\mathcal{H}$, $\mathcal{H}_1$, and $\mathcal{H}_2$ be Hilbert spaces and let $A\colon \mathcal{H}_1 \to \mathcal{H}$ and $B\colon \mathcal{H}_2 \to \mathcal{H}$ be bounded linear operators with $\operatorname{ran} A \subseteq \operatorname{ran} B$. Then $Q := B^\dagger A\colon \mathcal{H}_1 \to \mathcal{H}_2$ is a well-defined and bounded linear operator, where $B^\dagger$ denotes the Moore–Penrose pseudo-inverse of B. It is the unique operator that satisfies the conditions*

$$A = BQ, \qquad \ker Q = \ker A, \qquad \operatorname{ran} Q \subseteq \overline{\operatorname{ran} B^*}. \tag{A.1}$$

Remark A.2. In the original work of [9] only the existence of a bounded operator Q such that $A = BQ$ was shown. However, the construction of Q in the proof is identical to that of $B^\dagger$ (multiplied by A). This connection has been observed before by [1], Corollary 2.2 and Remark 2.3, where it was proven in the case of closed range operators, leaving the proof of the general case to the reader. Moreover, [9], Theorem 1, only treats the case $\mathcal{H} = \mathcal{H}_1 = \mathcal{H}_2$; the general case is mentioned as a remark at the end of his paper.

Further, we are going to use the following characterisations of cross-covariance operators due to [3], Theorem 1.

Theorem A.3. *Under the notation of Section 3, there exists a unique bounded linear operator $R_{VU}\colon \mathcal{G} \to \mathcal{H}$ with operator norm $\|R_{VU}\| \leq 1$ such that*

$$C_{VU} = C_V^{1/2} R_{VU} C_U^{1/2}, \qquad R_{VU} = P_{(\ker C_V)^\perp} R_{VU} P_{(\ker C_U)^\perp}. \tag{A.2}$$

Remark A.4. If $\mathcal{H} = \mathcal{G} = \mathbb{R}$, then R_{VU} coincides with the Pearson correlation coefficient.

This paper makes extensive use of the following two basic results.

Lemma A.5. *Let $A\colon \mathcal{H} \to \mathcal{G}$ be a trace-class operator such that $\operatorname{tr}(AB^*) = 0$ for any bounded operator $B \in \mathsf{L}(\mathcal{H}; \mathcal{G})$. Then $A = 0$.*

Proof. Choosing $B = A$ yields $\|A\|_{\mathsf{L}_2} = \operatorname{tr}(AA^*)^{1/2} = 0$, hence $A = 0$. □

Lemma A.6. *With the notation of Section 3, let $U' \in L^2(\Omega, \Sigma, \mathbb{P}; \mathcal{G})$ and $\gamma \in \mathsf{A}_V(\mathcal{H}; \mathcal{G})$. Then*

(a) $\langle U - \mu_U, U' \rangle_{L^2(\mathbb{P};\mathcal{G})} = \operatorname{tr}(\mathbb{C}\mathrm{ov}[U, U'])$;
(b) $\mathbb{C}\mathrm{ov}[\gamma \circ V, U] = \overline{\gamma} C_{VU}$ *and* $\mathbb{C}\mathrm{ov}[U, \gamma \circ V] = (\overline{\gamma} C_{VU})^*$.

If $\gamma \in \mathsf{A}(\mathcal{H}; \mathcal{G})$, then the last equation can be simplified to $\mathbb{C}\mathrm{ov}[U, \gamma \circ V] = C_{UV} \overline{\gamma}^*$.

Proof. Let $(e_j)_{j \in \mathcal{J}}$, $\mathcal{J} \subseteq \mathbb{N}$ be an orthonormal basis of $\mathcal{G}$. Then

$$\begin{aligned}
\langle U - \mu_U, U' \rangle_{L^2(\mathbb{P};\mathcal{G})} &= \langle U - \mu_U, U' - \mu_{U'} \rangle_{L^2(\mathbb{P};\mathcal{G})} \\
&= \mathbb{E}\big[\langle U - \mu_U, U' - \mu_{U'} \rangle_{\mathcal{G}}\big] \\
&= \sum_{j \in \mathcal{J}} \mathbb{E}\big[\langle e_j, U - \mu_U \rangle_{\mathcal{G}} \langle U' - \mu_{U'}, e_j \rangle_{\mathcal{G}}\big]
\end{aligned}$$

$$= \sum_{j \in \mathcal{J}} \langle e_j, C_{UU'} e_j \rangle_{\mathcal{G}}$$
$$= \operatorname{tr}\big[\mathbb{C}\mathrm{ov}\big[U, U'\big]\big],$$

proving (a). Since $V \in L^2(\mathbb{P}; \mathcal{H})$ and $\gamma \circ V \in L^2(\mathbb{P}; \mathcal{G})$, all covariance operators are well defined and so, for $g \in \mathcal{G}$,

$$\mathbb{C}\mathrm{ov}[\gamma V, U](g) = \mathbb{E}\big[\gamma(V - \mu_V)\langle U - \mu_U, g\rangle_{\mathcal{G}}\big] = \gamma \mathbb{E}\big[(V - \mu_V)\langle U - \mu_U, g\rangle_{\mathcal{G}}\big] = \gamma \mathbb{C}\mathrm{ov}[V, U](g),$$

proving (b). □

Funding

IK and TJS are supported in part by the Deutsche Forschungsgemeinschaft (DFG) through project TrU-2 "Demand modelling and control for e-commerce using RKHS transfer operator approaches" of the Excellence Cluster "MATH+ The Berlin Mathematics Research Centre" (EXC-2046/1, project 390685689). TJS is further supported by the DFG project 415980428. BS has been supported by the DFG project 389483880.

References

[1] Arias, M.L., Corach, G. and Gonzalez, M.C. (2008). Generalized inverses and Douglas equations. *Proc. Amer. Math. Soc.* **136** 3177–3183. MR2407082 https://doi.org/10.1090/S0002-9939-08-09298-8

[2] Aubin, J.-P. (2000). *Applied Functional Analysis*, 2nd ed. *Pure and Applied Mathematics* (*New York*). New York: Wiley Interscience. MR1782330 https://doi.org/10.1002/9781118032725

[3] Baker, C.R. (1973). Joint measures and cross-covariance operators. *Trans. Amer. Math. Soc.* **186** 273–289. MR0336795 https://doi.org/10.2307/1996566

[4] Brémaud, P. (2017). *Discrete Probability Models and Methods*: *Probability on Graphs and Trees*, *Markov Chains and Random Fields*, *Entropy and Coding*. *Probability Theory and Stochastic Modelling* **78**. Cham: Springer. MR3616988 https://doi.org/10.1007/978-3-319-43476-6

[5] Chatterji, S.D. (1960). Martingales of Banach-valued random variables. *Bull. Amer. Math. Soc.* **66** 395–398. MR0119242 https://doi.org/10.1090/S0002-9904-1960-10471-5

[6] Chilès, J.-P. and Delfiner, P. (2012). *Geostatistics*: *Modeling Spatial Uncertainty*, 2nd ed. *Wiley Series in Probability and Statistics*. Hoboken, NJ: Wiley. MR2850475 https://doi.org/10.1002/9781118136188

[7] Corach, G., Maestripieri, A. and Stojanoff, D. (2001). Oblique projections and Schur complements. *Acta Sci. Math.* (*Szeged*) **67** 337–356. MR1830147

[8] Diestel, J. and Uhl, J.J. Jr. (1977). *Vector Measures*. Providence, RI: Amer. Math. Soc. MR0453964

[9] Douglas, R.G. (1966). On majorization, factorization, and range inclusion of operators on Hilbert space. *Proc. Amer. Math. Soc.* **17** 413–415. MR0203464 https://doi.org/10.2307/2035178

[10] Dudley, R.M. (2002). *Real Analysis and Probability*. *Cambridge Studies in Advanced Mathematics* **74**. Cambridge: Cambridge Univ. Press. MR1932358 https://doi.org/10.1017/CBO9780511755347

[11] Engl, H.W., Hanke, M. and Neubauer, A. (1996). *Regularization of Inverse Problems*. *Mathematics and Its Applications* **375**. Dordrecht: Kluwer Academic. MR1408680

[12] Ernst, O.G., Sprungk, B. and Starkloff, H.-J. (2015). Analysis of the ensemble and polynomial chaos Kalman filters in Bayesian inverse problems. *SIAM/ASA J. Uncertain. Quantificat.* **3** 823–851. MR3400030 https://doi.org/10.1137/140981319

[13] Evensen, G. (2009). *Data Assimilation*: *The Ensemble Kalman Filter*, 2nd ed. Berlin: Springer. MR2555209 https://doi.org/10.1007/978-3-642-03711-5

[14] Fillmore, P.A. and Williams, J.P. (1971). On operator ranges. *Adv. Math.* **7** 254–281. MR0293441 https://doi.org/10.1016/S0001-8708(71)80006-3
[15] Folland, G.B. (1999). *Real Analysis*: *Modern Techniques and Their Applications*, 2nd ed. *Pure and Applied Mathematics* (*New York*). New York: Wiley. MR1681462
[16] Fukumizu, K., Song, L. and Gretton, A. (2013). Kernel Bayes' rule: Bayesian inference with positive definite kernels. *J. Mach. Learn. Res.* **14** 3753–3783. MR3159407
[17] Goldstein, M. (1999). Bayes linear analysis. In *Encyclopaedia of Statistical Sciences* (S. Kotz, B.C. Read, N. Balakrishnan, B. Vidakovic and N.L. Johnson, eds.) 29–34. Chichester: Wiley.
[18] Goldstein, M. and Wooff, D. (2007). *Bayes Linear Statistics*: *Theory and Methods*. *Wiley Series in Probability and Statistics*. Chichester: Wiley. MR2335584 https://doi.org/10.1002/9780470065662
[19] Hairer, M., Stuart, A.M., Voss, J. and Wiberg, P. (2005). Analysis of SPDEs arising in path sampling. I. The Gaussian case. *Commun. Math. Sci.* **3** 587–603. MR2188686
[20] Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning*: *Data Mining*, *Inference*, *and Prediction*, 2nd ed. *Springer Series in Statistics*. New York: Springer. MR2722294 https://doi.org/10.1007/978-0-387-84858-7
[21] Kallenberg, O. (2006). *Foundations of Modern Probability*. *Probability and Its Applications* (*New York*). New York: Springer. MR1464694
[22] Klebanov, I., Schuster, I. and Sullivan, T.J. (2020). A rigorous theory of conditional mean embeddings. *SIAM J. Math. Data Sci.* **2** 583–606. MR4121886 https://doi.org/10.1137/19M1305069
[23] Klenke, A. (2013). *Wahrscheinlichkeitstheorie*, 3rd ed. Berlin: Springer. https://doi.org/10.1007/978-3-642-36018-3
[24] Klus, S., Husic, B.E., Mollenhauer, M. and Noé, F. (2019). Kernel methods for detecting coherent structures in dynamical data. *Chaos* **29** 123112, 15. MR4041090 https://doi.org/10.1063/1.5100267
[25] Klus, S., Nüske, F., Koltai, P., Wu, H., Kevrekidis, I., Schütte, C. and Noé, F. (2018). Data-driven model reduction and transfer operator approximation. *J. Nonlinear Sci.* **28** 985–1010. MR3800253 https://doi.org/10.1007/s00332-017-9437-7
[26] Klus, S., Schuster, I. and Muandet, K. (2020). Eigendecompositions of transfer operators in reproducing kernel Hilbert spaces. *J. Nonlinear Sci.* **30** 283–315. MR4054854 https://doi.org/10.1007/s00332-019-09574-z
[27] Mandelbaum, A. (1984). Linear estimators and measurable linear transformations on a Hilbert space. *Z. Wahrsch. Verw. Gebiete* **65** 385–397. MR0731228 https://doi.org/10.1007/BF00533743
[28] Meise, R. and Vogt, D. (1997). *Introduction to Functional Analysis*. *Oxford Graduate Texts in Mathematics* **2**. New York: The Clarendon Press, Oxford Univ. Press. MR1483073
[29] Owhadi, H. and Scovel, C. (2018). Conditioning Gaussian measure on Hilbert space. *J. Math. Stat. Anal.* **1** 1–15.
[30] Sazonov, V. (1958). On characteristic functionals. *Teor. Veroyatn. Primen.* **3** 201–205. MR0098423
[31] Schillings, C. and Stuart, A.M. (2017). Analysis of the ensemble Kalman filter for inverse problems. *SIAM J. Numer. Anal.* **55** 1264–1290. MR3654885 https://doi.org/10.1137/16M105959X
[32] Schwantes, C.R. and Pande, V.S. (2015). Modeling molecular kinetics with tICA and the kernel trick. *J. Chem. Theory Comput.* **11** 600–608. https://doi.org/10.1021/ct5007357
[33] Song, L., Huang, J., Smola, A. and Fukumizu, K. (2009). Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *Proceedings of the* 26*th Annual International Conference on Machine Learning* 961–968. https://doi.org/10.1145/1553374.1553497
[34] Stein, M.L. (1999). *Interpolation of Spatial Data*: *Some Theory for Kriging*. *Springer Series in Statistics*. New York: Springer. MR1697409 https://doi.org/10.1007/978-1-4612-1494-6
[35] Steinwart, I. and Christmann, A. (2008). *Support Vector Machines*. *Information Science and Statistics*. New York: Springer. MR2450103
[36] Tarieladze, V. and Vakhania, N. (2007). Disintegration of Gaussian measures and average-case optimal algorithms. *J. Complexity* **23** 851–866. MR2371996 https://doi.org/10.1016/j.jco.2007.04.005

Received September 2020 and revised December 2020